МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
СХІДНОУКРАЇНСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ ІМ. В. ДАЛЯ
ФАКУЛЬТЕТ ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ ТА ЕЛЕКТРОНІКИ
КАФЕДРА КОМП'ЮТЕРНИХ НАУК ТА ІНЖЕНЕРІЇ

До захисту допускається
Завідувач кафедри
_____ Скарга-Бандурова І.С.
«____»_____ 20__р.

МАГІСТЕРСЬКА РОБОТА

НА ТЕМУ:

Методи та інформаційна технологія проектування автоматизованих систем обробки інформації з соціальних мереж

Освітньо-кваліфікаційний рівень "Магістр"
Спеціальність 122 – "Комп'ютерні науки та інформаційні технології" (освітня програма – "Інформаційні технології проектування")

Науковий керівник роботи:	_____ (підпис)	І.С. Скарга-Бандурова (ініціали, прізвище)
Консультант з охорони праці:	_____ (підпис)	Я.О. Критська (ініціали, прізвище)
Студент:	_____ (підпис)	І.М. Баєв (ініціали, прізвище)
Група:		ІТП-16дм

Сєвєродонецьк 2018

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
СХІДНОУКРАЇНСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ВОЛОДИМИРА ДАЛЯ

Факультет Інформаційних технологій та електроніки
Кафедра Комп'ютерних наук та інженерії
Освітньо-кваліфікаційний рівень магістр
Напрям підготовки _____
(шифр і назва)
Спеціальність 122 – "Комп'ютерні науки та інформаційні технології"
(шифр і назва)

ЗАТВЕРДЖУЮ:

Завідувач кафедри _____
_____ І.С. Скарга-Бандурова
«_____» _____20____р.

ЗАВДАННЯ
НА МАГІСТЕРСЬКУ РОБОТУ СТУДЕНТУ

Баєву Іллі Михайловичу
(прізвище, ім'я,по батькові)

1. Тема роботи Методи та інформаційна технологія проектування автоматизованих систем обробки інформації з соціальних мереж

керівник проекту (роботи) Скарга-Бандурова І.С
(прізвище, ім'я, по батькові, науковий ступінь, вчене звання)

затверджені наказом вищого навчального закладу від " 18 " 10 2017 року № 207/48

2. Строк подання студентом проекту (роботи) 12.01.2018

3. Вихідні дані до проекту (роботи) матеріали науково-дослідницької практики

4.Зміст розрахунково-пояснювальної записки (перелік питань, які потрібно розробити)

Аналіз програмних та інструментальних засобів для отримання інформації з соціальних мереж. Аналіз математичних моделей і методів вирішення задачі виявлення тональності у тексті. Аналіз методології побудови моделей класифікатора. Виявлення проблем при роботі з інформацією з соціальних мереж. Розробка допоміжного ПЗ. Проведення експериментів. Охорона праці

5. Перелік графічного матеріалу (з точним зазначенням обов'язкових креслеників)

електронні плакати

6. Консультанти розділів проекту (роботи)

Розділ	Прізвище, ініціали та посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв
Охорона праці	Критська Яна Олександрівна		

7. Дата видачі завдання 18.10.2017

Керівник _____
(підпис)

Завдання прийняв до виконання _____
(підпис)

КАЛЕНДАРНИЙ ПЛАН

№ п/п	Найменування етапів дипломного проекту (роботи)	Термін виконання етапів проекту (роботи)	Примітка
1	Аналіз стану питання у науковій літературі. Визначення вимог до роботи.	18.10.2017 – 16.11.2017	
2	Аналіз програмних та інструментальних засобів для роботи з даними з соціальних мереж.	17.11.2017 – 22.11.2017	
3	Аналіз математичних моделей і методів вирішення задачі виявлення тональності у тексті. Проведення початкового експерименту.	23.11.2017 – 28.11.2017	
4	Опис методології побудови тренувальної та прогнозуючої моделей класифікатора на прикладі цих моделей у сервісі MS Azure ML Studio.	29.11.2017 – 07.12.2017	
5	Огляд методів боротьби з незбалансованими даними та засобів оцінки моделей класифікації.	08.12.2017 – 12.12.2017	
6	Розробка додаткового ПЗ для використання міри дельта TF-IDF для розрахунку ваг слів.	13.12.2017 – 20.12.2017	
7	Розробка заходів з охорони праці.	21.12.2017 – 24.12.2017	
8	Оформлення пояснювальної записки і графічного матеріалу.	25.12.2017 – 06.01.2018	
9	Підготовка та подання магістерської роботи до захисту	07.01.2018 – 12.01.2018	

Студент _____ Баєв І.М.
(підпис) (прізвище та ініціали)

Науковий керівник _____ Скарга-Бандурова І.С.
(підпис) (прізвище та ініціали)

АНОТАЦІЯ

Баєв І.М. Методи та інформаційна технологія проектування автоматизованих систем обробки інформації с соціальних мереж.

Розглянуті програмні та інструментальні засоби отримання інформації з соціальних мереж на прикладі Twitter. Проведено аналіз математичних моделей і методів вирішення задачі виявлення тональності в тексті. Проведений базовий експеримент емоційної класифікації виявив проблему незбалансованості даних. Викладено загальну методику побудови моделей класифікаторів. Розглянуті методи боротьби з незбалансованими даними. Розроблено допоміжне ПЗ.

Ключові слова: емоційна класифікація, класифікатор, програмне забезпечення, незбалансовані дані, векторно-просторова модель

АНОТАЦИЯ

Баев И.М. Методы и информационные технологии проектирования автоматизированных систем обработки информации с социальных сетей.

Рассмотрены программные и инструментальные средства получения информации из социальных сетей на примере Twitter. Проведен анализ математических моделей и методов решения задачи обнаружения тональности в тексте. Проведенный базовый эксперимент эмоциональной классификации обнаружил проблему несбалансированности данных. Изложены общую методику построения моделей классификаторов. Рассмотрены методы борьбы с несбалансированными данными. Разработано вспомогательное ПО.

Ключевые слова: эмоциональная классификация, классификатор, программное обеспечение, несбалансированные данные, векторно-пространственная модель

ABSTRACT

Baev I.M. Methods and information technology of designing automated information processing systems from social networks.

Software and tools for obtaining information from social networks using an example of Twitter are considered. The analysis of mathematical models and methods of solving the problem of sentiment analysis is carried out. The basic experiment of emotional classification has revealed the problem of imbalance data. The general methodology for constructing classifier models is described. Methods of combating unbalanced data are considered. Auxiliary software is developed.

Keywords: emotional classification, classifier, software, unbalanced data, vector space model

ЗМІСТ

ВСТУП

Обґрунтування вибору теми досліджень. Нещодавно межа між звичайними користувачами і тими, хто створює інформацію в Інтернеті зникла: замість статичних сторінок у всіх користувачів з'явилася можливість публікувати свою інформацію. Зараз ми спостерігаємо величезну кількість видів створюваних матеріалів: це може бути запис в блозі або на форумі, фотографія або відеозапис на відповідному ресурсі, відгук в Інтернет-магазині, «статус» у соціальній мережі і багато іншого. Простота розміщення текстів чи іншої інформації в одному місці підштовхнуло розробників створювати сайти, де люди можуть вільно викладати свої думки про будь що: фільми, книжки, людей тощо. Такі сайти грають провідну роль у житті сучасної людини: перед тим, як купити той чи інший товар, покупець читає багато коментарів та відгуків та робить висновок купляти йому цей товар чи шукати інший. Згодом текстів стало так багато, що обробити їх людині за короткий проміжок часу стало просто неможливим. Саме така ситуація стала причиною появи задачі аналізу думок. З’явилась потрібність в системах, що змогли б автоматично шукати і класифікувати думки.

Аналіз думок – один із напрямів галузі обробки текстів на природних мовах. Спочатку в якості даних, що досліджувались використовувалися великі записи, що складаються з декількох речень, в яких явно простежувався зв'язок. Якщо раніше для того, щоб виявити думку про яке-небудь питання треба було проводити опитування, то тепер, з розвитком соціальних мереж та появою в них коментарів, «статусів» і коротких повідомлень, користувацький контент став менш ємним, але при цьому більш суб'єктивним і перетворився в нескінченний потік інформації, який треба виявити і оцінити. Яскравим прикладом цього є сервіс мікроблогів Twitter За допомогою цього сервісу користувачі поширюють свої погляди на актуальні новини, політику, економіку, розповідають про свій настрій.

Саме про Twitter і буде надалі іти мова.

Головною особливістю цього сервісу є те, що користувач має викласти свою думку лише у 140 символах. Кожне повідомлення (твіт), яке публікується користувачем в Twitter, можуть побачити його підписники – люди, які пов'язані з ним в цій соціальній мережі. Підписники (або інакше читачі) можуть бути як односторонніми, так і взаємними. Якщо людина побачила, твіт, що її зацікавив і розмістила його на своїй сторінці, то кажуть, що вона ретвітнула запис іншого користувача. Є й інший вид взаємодії користувацької інформації: згадки. Якщо читач захотів відповісти на будь-який твіт, він

це робить, вставивши в початок свого повідомлення псевдонім автора в Twitter за допомогою символу @ (@username), тим самим, згадуючи його. У цьому випадку відповідь побачать тільки ті, хто читає обох дискутуючих користувачів.

Основний спосіб представлення твітів – це представлення у вигляді потоку. Коли користувач входить на сайт, то перше, що він бачить це повідомлення від усіх акаунтів що він «читає», відсортованих в хронологічному порядку. Також він може перейти до будь-якого акаунта і «прочитати» тільки його повідомлення. За допомогою ретвітів інформація поширюється дуже швидко.

Також користувач може виконати пошук повідомлень за хештегом. Хештег – спеціальне слово перед яким є символ #. Наявність хештега показує, що інформація в твіті має відношення до цього спеціального слова. Пошук за хештегом враховує записи усіх користувачів, тому кількість отриманої інформації дуже велика і користувач не може фізично її обробити. Така ж сама ситуація і з пошуком за ключовими словами.

Розглянемо приклад застосування аналізу думок в Twitter. Коли велика компанія випускає новий продукт, то вона публікує запис про це. Читачі твітер-акаунта цієї компанії бачать повідомлення про новий продукт і, по-перше, самі про неї дізнаються, по-друге, можуть ретвітнути її для своїх підписників, і до аудиторії новини приєднаються інші користувачі, по-третє, можуть прокоментувати і показати тим самим своє ставлення до події. На всіх етапах поширення інформації про продукт, компанії важливо, яке емоційне забарвлення вона несе. Тут вже починається дослідження цього забарвлення не серед коментарів до твітів і ретвітів конкретного запису, а в цілому серед текстів, які мають відношення до цільового об'єкту.

Для вирішення задачі виявлення емоційного забарвлення використовують різні методи і моделі. У загальному випадку: тренують модель класифікації на вже емоційно-розмічених текстах та використовують її для нових даних. Але після класифікації може виникнути проблема незбалансованих даних [7].

Тому обґрунтованою є тема магістерської роботи, у якій вирішується **науково-прикладне** завдання розроблення моделей і методу інформаційної технології поліпшення результатів емоційної класифікації в умовах небалансованих даних.

Об’єкт дослідження – процеси обробки текстів на природних мовах.

Предмет дослідження – моделі та методи, що використовуються для аналізу текстових даних, отриманих з соціальних мереж.

Мета і завдання дослідження. Метою дослідження є підвищення ефективності роботи з даними з соціальних мереж та підвищення точності емоційної класифікації за

рахунок впровадження моделей класифікації, здатних оперувати наборами незбалансованих даних.

Для досягнення мети дослідження необхідно вирішити такі **завдання**:

— виявлення проблем при роботі з даними з соціальних мереж;

— аналіз методів класифікації тексту;

— розроблення моделей класифікації;

— проектування інформаційної технології та розробка допоміжних програмних засобів.

Методи дослідження. Проведені в роботі дослідження основані на технології sentiment mining (аналіз думок) для пошуку залежностей між текстом та його емоційним забарвленням, методах машинного навчання для побудови класифікаційної моделі, методах боротьби з незбалансованими даними, зокрема міра розрахунку ваг слів у векторно-просторовій моделі дельта TF-IDF, яка використовувалась при розробці додатку, який і обчислює ці ваги.

Наукова новизна отриманих результатів:

1. Удосконалено тренувальну модель класифікатора шляхом застосування методів маніпулювання з незбалансованими даними, що дозволяє покращити роботу моделі обробки текстів на природній мові.

2. Дістали подальшого розвитку моделі прогнозування з використанням моделі дельта TF-IDF, що дозволяє підвищити точність класифікації.

Особистий внесок здобувача полягає у розроблені нових моделей та інструментальних засобів, що дозволяють вирішити поставлені задачі. Усі основні результати отримані автором особисто.

Апробація матеріалів дисертації. Основні положення, ідеї, висновки магістерської роботи доповідалися та обговорювалися на всеукраїнській конференції «ІТ-ІДЕЯ» 2016 та 2017 років, а також на другій міжнародній конференції «Theoretical and Applied Computer Science and Information Technologies» (м. Сєвєродонецьк, 2017 р.).

Практичне значення отриманих результатів полягає у тому, що основні наукові положення дисертації реалізовані у виді розрахункових моделей та програмних засобів, які утворюють прикладну інформаційну технологію проектування автоматизованих систем обробки інформації з соціальних мереж.

Розроблено моделі з урахуванням методів, які сприяють вирішенню проблеми незбалансованості даних, а також програмне забезпечення, яке реалізовує підхід розрахунку ваг слів або словосполучень дельта TF-IDF у векторно-просторовій моделі.

Використання цих моделей в інформаційній технології проектування автоматизованих систем обробки інформації з соціальних мереж дозволило покращити результати емоційного прогнозування класифікатора.

Публікації. За темою магістерської роботи з викладенням її основних результатів опубліковано 3 наукових праці серед яких тези 2 доповідей на всеукраїнській конференції IT-ідея 2016-17 рр. та тези доповіді на міжнародній конференції Theoretical and Applied Computer Science and Information Technology: Proceedings of the II International Conference 2017.

Структура та обсяг дисертації. Дисертація складається зі вступу, чотирьох розділів, висновків, списку використаних джерел і додатків. Загальний обсяг дисертації складає 128 сторінок, з яких анотація на 1 сторінці, зміст на 3 сторінках, вступ на 3 сторінках, основний текст на 83 сторінках, висновки на 1 сторінці, список використаних джерел із 66 найменувань на 6 сторінках, додатки на 26 сторінках. Робота містить 12 таблиць та 38 рисунків.

1 АНАЛІЗ МЕТОДІВ І ЗАСОБІВ ВИЯВЛЕННЯ ЕМОЦІЙ У ТЕКСТІ З СОЦІАЛЬНИХ МЕРЕЖ. ПОСТАНОВКА ЗАДАЧІ ДОСЛІДЖЕНЬ

У розділі проаналізовані програмні та інструментальні засоби для взаємодії з Twitter, математичні моделі і методи вирішення задачі виявлення тональності у тексті, проведений початковий експеримент для виявлення проблем, поставлена наукова задача та обґрунтована методика досліджень.

1.1 Аналіз програмних та інструментальних засобів

1.1.1 Twitter API

Для взаємодії з Twitter використовуються спеціальні інтерфейси – API.

Прикладний програмний інтерфейс, або API – набір визначень взаємодії різнотипного програмного забезпечення.[8].

Існує три API, що дозволяють отримати дані з Twitter:

— Twitter Search API

— Twitter Streaming API

— Twitter Firehose/PowerTrack

— Twitter Premium APIs

У таблиці 1.1 показані основні характеристики API.

Таблиця 1.1 – Основні характеристики API

Назва	**Обмеження**	**Необхідність стороннього ПО**	**Ціна**
Twitter Search API	Так	Так	Безкоштовно
Twitter Streaming API	Так	Так	Безкоштовно
Twitter Firehose/PowerTrack	Ні	Так	$99-2000
Twitter Premium APIs	Так	Так	$149/місяць

Нижче характеристики описані більш детально.

1.1.1.1 Twitter Search API

Twitter Search API є частиною API REST для Twitter. Цей API повертає вибірку останніх твітів, опублікованих за останні 7 днів [9].

Через Twitter Search API користувач запитує твіти, які відповідають певним критеріям. Цими критеріями можуть бути ключові слова, імена користувачів, хештеги, назви місць та інше. Прикладом може слугувати звичайний пошук на самому сайті Twitter.

При використанні цього API розробники стикаються з деякими обмеженнями. Для окремого користувача максимальна кількість твітів – останні 3200 незалежно від критеріїв запиту. За допомогою певного ключового слова – 5000 твітів [10]. Також розробник обмежений певною кількістю запитів, які можна зробити за певний проміжок часу. Цей інтерфейс повертає користувачу ті твіти, які вже «відбулися», тобто твіти з не найближчою датою опублікування. Застосовувати цей API можна безкоштовно.

1.1.1.2 Twitter Streaming API

На відміну від Search API Streaming API повертає користувачу твіти, які відбуваються практично в реальному часі. За допомогою цього інтерфейсу можна отримувати інформацію не тільки за певними критеріями як у Search API, але і від багатьох користувачів без застосовування критеріїв.

Основним недоліком Streaming API є те, що цей інтерфейс надає не всі твітіи, які відбуваються. Фактичний відсоток отриманих твітів сильно залежить від критеріїв запитів користувача і поточного трафіку. Тобто користувачі можуть очікувати отримання від 1% до більш ніж 40% твітів [10]. Причиною цього слугує те, що Twitter не має поточної інфраструктури для підтримки інтерфейсу. Застосовувати цей API можна безкоштовно.

1.1.1.3 Twitter Firehose/PowerTrack

Ще одним способом доступу до даних є Twitter Firehose. Twitter Firehose насправді дуже схожий на Twitter Streaming API, оскільки він повертає дані до кінцевого користувача майже в режимі реального часу, але Twitter Firehose гарантує доставку 100% твітів, які відповідають критеріям запиту користувача.

Головною відмінністю між Streaming API і Twitter Firehose є те, що ви гарантовано отримуєте 100% твітів, і це не безкоштовно (доступ до цих даних дають провайдери, які

тісно пов’язані з Twitter). API Twitter Streaming можна використовувати, але він надає обмежені результати (і обмежене використання ліцензій даних). Доступ до Twitter Firehose знімає багато обмежень щодо використання, але це досить дорого для доступу до всіх твітів.

Наразі Twitter Firehose називається PowerTrack API. PowerTrack API надає клієнтам можливість фільтрувати дані і отримувати лише ті, в яких вони зацікавляться. Це виконується за допомогою мови фільтрації PowerTrack. Використання правил PowerTrack для фільтрування даних гарантує, що клієнти отримуватимуть усі дані, які вони потребують [9]. Доступ до API пре надає компанія Gnip, що надає дані з соціальних мереж. Була куплена Twitter у 2014 році.

Щоб отримати можливість роботи з цим API, треба завести особливий (enterprise) аккаунт у Twitter. Нажаль на сайті Twitter немає ціни за користування.

Також є сайт dicovertext, який теж дозволяє працювати з PowerTrack API. Він пропонує ціни від $99 до $2000 за місяць [11]. Також можливий пробний варіант, який дозволяє отримати 500 твітів. Для збільшення їх кількості пропонують заплатити $5 за 10000 твітів [12].

Отже для роботи з Twitter доцільно використовувати Streaming та Search API.

1.1.1.4 Twitter Premium APIs

Twitter обіцяє, що навесні 2018 року випустить преміальні API. Ці інтерфейси мають забезпечити надійність корпоративних API для екосистеми розробників. Наразі вони працюють в режимі публічної бети, та пропонують преміальний Search API, який надає доступ до даних за останні 30 днів. Крім цього API мають:

— більшу кількість твітів, що повертаються за один запит;
— більш висока загальна кількість запитів;
— більш складі запити;
— нові метадані, наприклад розширені URL.

1.1.2 Сторонні ПЗ для Twitter API

Для роботи за API треба використовувати стороннє ПЗ. Розглянемо їх на прикладі Microsoft Visual Studio та Microsoft SQL Server.

1.1.2.1 Microsoft Visual Studio

Microsoft Visual Studio – інтегроване середовище розробки програмного забезпечення та ряд інших інструментальних засобів, розроблене корпорацією Microsoft. Нажаль, воно не є безкоштовним та остання професійна версія коштує $499. Але існує і Community версія для учнів, розробників відкритого ПЗ та окремих розробників, яка є безкоштовною [13].

Для роботи з даними з Twitter краще використовувати об’єктно-орієнтовну мову програмування, таку як C#.

Для взаємодії між Twitter API та Visual Studio можна застосовувати відкриті бібліотеки:

— LINQ2Twitter;

— Spring.NET Social Extension for Twitter;

— TweetSharp;

— Tweetinvi;

— Crafted.Twitter.

Вони дають можливість отримувати данні і за допомогою Search API, і за допомогою Streaming API.

1.1.2.2 Microsoft SQL Server

Щоб не створювати навантаження на апаратне забезпечення та, щоб не зменшувалась швидкість роботи створюваної програми треба застосовувати базу даних для накопичення та подальшої обробки даних.

Microsoft SQL Server - система управління базами даних (РСУБД), розроблена корпорацією Microsoft. Основна мова запитів - Transact-SQL, створена спільно Microsoft та Sybase. Transact-SQL є реалізацією стандарту ANSI / ISO з структурованої мови запитів (SQL) з розширеннями. Використовується для роботи з базами даних розміром від персональних до великих баз даних масштабу підприємства; конкурує з іншими СУБД в цьому сегменті ринку. Остання версія для розробників є безкоштовною.

Microsoft SQL Server призначена виключно для підтримки систем, що працюють в середовищі клієнт-сервер. Вона підтримує широкий спектр засобів розробки і максимально простий в інтеграції з додатками, що працюють на ПК.

Ця СУБД дозволяє шифрувати базу даних, файли даних або файли журналів, не вносячи в додатки ніяких змін. Завдяки цьому стає можливим пошук в зашифрованих

даних як за діапазоном, так і з нечіткими критеріями, а також пошук в захищених даних, отриманих від неавторизованих користувачів.

Вбудовані запити LINQ (Language Integrated Query) дозволяють розробникам замість використання SQL-запитів звертатися до даних з програм на керованих мовах, наприклад, C# або VB.NET. У SQL Server є можливість використання LINQ для звернення безпосередньо до таблиць і полів SQL Server.

SQL Server - велика розширювана платформа для організації сховищ даних, яка дозволяє швидше і ефективніше інтегрувати інформацію в сховища, а також керувати зростаючими її обсягами, надаючи всім користувачам відомості, необхідні для кращого розуміння цієї інформації.

В СУБД вбудована підтримка .NET Framework. Завдяки цьому, збережені процедури БД можуть бути написані на будь-якій мові платформи .NET, використовуючи повний набір бібліотек, доступних для .NET Framework, включаючи Common Type System.

1.1.3 Microsoft Azure Machine Learning Studio

Azure Machine Learning – хмарний сервіс машинного навчання, розробником якого виступає Microsoft. Цей сервіс має 2 види: вільний та стандарт. Перший – безкоштовний, але має деякі обмеження, такі як кількість модулів у експерименті, максимальне місце для зберігання, тощо. Другий коштує $99.99 на місяць на один акаунт [14].

Для того, щоб виконувати задачі аналітики у цьому сервісі достатньо зробити наступне [15]:

— Завантажити або імпортувати дані;

— Збудувати та валідувати модель;

— Створити веб-сервіс, який використає ваші моделі для передбачень в реальному часі.

Azure Machine Learning Studio має багато модулів для створення моделей (рис. 1.1): трансформація даних, фільтри автоматичного вибору функцій (Feature Selection), статистичні функції, машинне навчання, модулі для виконання скриптів на мовах R та Python та ін.

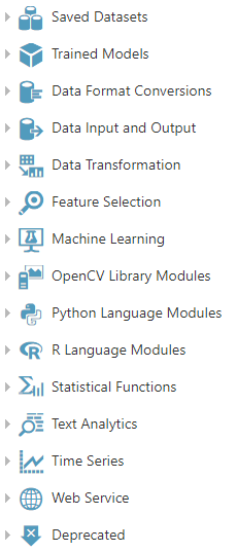

Рисунок 1.1 – Модулі у Machine Learning Studio

Як сказано в [15] Azure ML пропонує багато алгоритмів для класифікації задач, включно Multiclass та Two-Class Decision Forests, Decision Jungles (розроблено Microsoft Research), Logistic Regression, Neural Networks, а також Two-Class Averages Perceptrons, Bayes Point Machine, Boosted Decision Trees та Support Vector Machines (SVM).

Також у цьому сервісі присутня задача Named Entity Recognition, яка дозволяє обробляти вхідний текст (твіти, чи щось інше) та витягувати іменовані терміни, класифікуючи їх як організації, людей, тощо.

1.2 Аналіз математичних моделей і методів вирішення задачі виявлення тональності у тексті

1.2.1 Knowledge discovery in databases

В основі методів вирішення задач емоційної класифікації лежить процес Knowledge discovery in Databases.

Knowledge Discovery in Databases (KDD) - це процес пошуку корисних знань в «сирих» даних. KDD включає в себе питання: підготовки даних, вибору інформативних ознак, очищення даних, застосування методів Data Mining (DM), постобробки даних та інтерпретації отриманих результатів. Цей процес показано на рисунку 1.2.

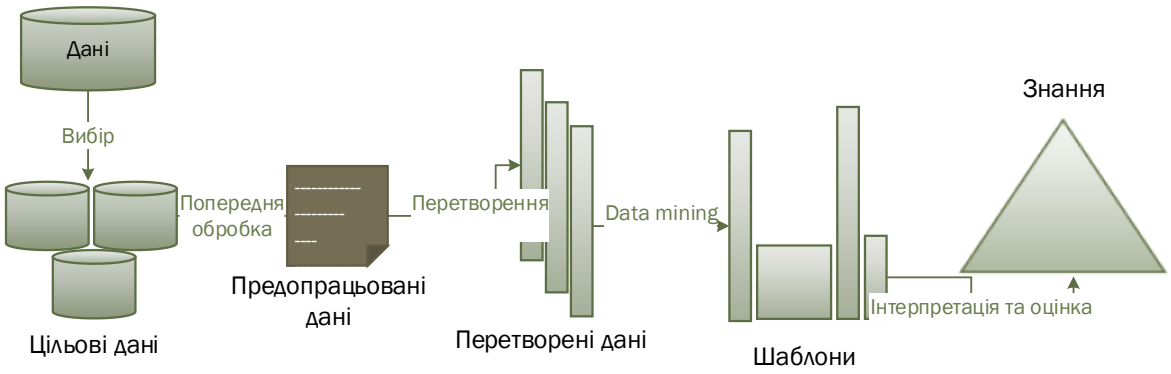

Рисунок 1.2 – Процес Knowledge Discovery in Databases (KDD).

Процес KDD бере необроблені (сирі) дані в якості вхідних даних і знаходить статистично значущі шаблони в цих даних (тобто знання) в якості висновку. З необроблених даних вибирається підмножина для обробки, яка позначається як *цільові дані*. Ці дані попередньо повинні бути оброблені, щоб підготувати їх до аналізу з використанням методів Data Mining (інтелектуального аналізу даних).

Для того щоб ефективно застосовувати методи Data Mining, слід звернути увагу на питання попередньої обробки даних. Дані можуть містити пропуски, шуми, аномальні значення і т.д. Також вони повинні бути якісні і коректні з точки зору використовуваного методу Data mining. Тому етапу KDD, який полягає в попередній обробці даних слід приділяти значну увагу.

Наступним кроком є застосування інтелектуального аналізу даних на попередньо оброблених (і перетворених) даних для вилучення цікавих шаблонів. Ці шаблони оцінюються для забезпечення їх достовірності та обґрунтованості та інтерпретації для забезпечення розуміння даних.

В області роботи з соціальними мережами необроблені дані – це контент, створюваний окремими людьми, а знання охоплюють цікаві закономірності, які спостерігаються в цих даних. Наприклад, для онлайн-продавця книг необроблені дані - це список книг, які купують люди, а цікава модель може описувати книги, які люди часто купують. У нашому випадку даними виступають твіти користувачів.

В процесі KDD дані представляються в табличному форматі. У соціальних мережах люди генерують багато типів НЕ табличних даних, таких як текст, голос чи відео. Ці типи даних спочатку треба перетворити в табличні дані, а лише потім обробляти з використанням алгоритмів інтелектуального аналізу даних.

Для перетворення тексту в табличний формат, можна використовувати процес, званий векторизацією, або вилученням ознак. Існує безліч методів векторизації. Відомим методом векторизації є векторно-просторова модель.

1.2.2 Векторно-просторова модель

Векторна модель – алгебраїчне представлення колекції документів векторами одного спільного для всієї колекції векторного простору. Векторна модель є основою для вирішення багатьох завдань інформаційного пошуку: класифікація документів, кластеризація документів ті ін [17].

Документ у векторній моделі розглядається як невпорядкована множина термів. Терм – слова та словосполучення з яких складається текст. Саме слово в загальному випадку є уніграмою (N-грама у якої N дорівнює 1).

N-грама — послідовність з n елементів. З семантичної точки зору, це може бути послідовність звуків, складів, слів або букв. На практиці частіше зустрічається N-грами як ряд слів. Послідовність з двох послідовних елементів часто називають біграм, послідовність з трьох елементів називається триграма.

Корпус або колекція текстів – корпус даних одиницями, якого є тексти або їх достатньо значні фрагменти.

Для визначення ваги терма в документі використовують такі функції зважування:

— TF (частота терма) – вага визначається як функція від кількості входження терма в документі;

— TF-IDF (частота терма- зворотна частота документа) – вага визначається як множення функції від кількості входжень терма у документ і функції від величини, зворотній кількості документів в колекції;

— булева вага – дорівнює 1, якщо терм зустрічається в документі та 0, якщо ні.

Мета векторної моделі полягає в тому, щоб перетворити текстові документи в вектори (ознаки).

Вектор документа можна представити у вигляді:

$$d_j = (w_{1j}, w_{2j}, \dots, w_{ij}), \tag{1.1}$$

де d_j – векторне представлення j-ого документа;

w_{ij} – вага i-ого терма в j-ому документі;

i – загальна кількість різних термів у всіх документах колекції.

У вигляді таблиці вектор можна представити так:

Таблиця 1.2 – Векторно-просторова модель

	Документ 1	Документ 2	…	Документ j
Терм 1	w_{11}	w_{12}	…	w_{1j}
Терм 2	w_{21}	w_{22}	…	w_{2j}
…	…	…	…	…
Терм i	w_{i1}	w_{i2}	…	w_{ij}

Також слід зазначити, що термів може бути дуже багато, то включати до вектору можна лише ті терми, які з’явилися у документі.

У нашому випадку документ дорівнює твіту.

Узагальнений підхід розрахунку ваги терма полягає у використанні так званої зваженої схеми «частота з якою термін зустрічається – зворотна частота документа» або Term Frequency - Inverse Document Frequency (TF-IDF). Показник w_{ij} за схемою TF-IDF розраховується таким чином:

$$w_{ij} = TF_{ij} \times IDF_i \tag{1.2}$$

TF розраховується як відношення числа входжень слова до загальної кількості слів у документі:

$$\mathrm{TF}(t,d) = \frac{n_t}{\sum_k n_k}, \tag{1.3}$$

де n_t – число входжень ознаки/терма t у документ;

знаменник – загальна кількість слів.

IDF – обраховується як інверсія частоти, з якою слово зустрічається у документах колекції:

$$IDF(t,D) = log\frac{|D|}{|\{d_i \in D | t \in d_i\}|}, \tag{1.4}$$

де |D| - число документів у корпусі;

$|\{d_i \in D \mid t \in d_i\}|$ - число документів з корпусу D, в яких зустрічається ознака/терм t.

Вибір основи логарифма в формулі не має значення, оскільки зміна основи призводить до зміни ваги кожного слова на постійний множник, а вагове співвідношення не зміниться.

Підхід TF-IDF надає більш високі ваги словами, які менше часто зустрічаються в документах, і в той же час мають більш високі частоти в документі, який вони використовують. Це гарантує, що слова з високими значеннями TF-IDF можуть використовуватися в якості репрезентативних прикладів документів, до яких вони належать, а також, що стоп-словам, таким як «the», «і», «або» які є загальними для всіх документів, буде присвоєна менша вага. Але для аналізу тональності тексту цей підхід не дає хороших результатів, тому що для аналізу тональності не настільки важливі часто повторювані слова. Тому частіше використовують булеву вагу.

Однак, існують методи, які використовують вагу слів та дають більш кращі результати при класифікації тональності, наприклад дельта TF-IDF [22]. Ідея цього метода полягає у тому, щоб дати більшу вагу словам, які мають не нейтральну тональність. Вага слова обраховується так:

$$w_{i,j} = F_{i,j} \cdot \log\left(\frac{|Neg| \cdot Pos_i}{|Pos| \cdot Neg_i}\right), \quad (1.5)$$

де $w_{i,j}$ – вага слова і у документі j;

$F_{i,j}$ – частота слова і у документі j;

|Neg| - кількість документів з негативною тональністю;

Pos_i – кількість позитивних документів де зустрічається ознака/терм і;

|Pos| - кількість документів з позитивною тональністю;

Neg_i - кількість негативних документів де зустрічається ознака/терм і.

Після векторизації документи перетворюються на вектори та до них вже можна застосовувати загальні алгоритми інтелектуального аналізу даних, проте, для забезпечення якісних результатів, необхідно перевірити якість даних, провести їх попередню обробку і перетворення.

Окрім векторизації, яка основана на N-грамах, що описана вище, також можна застосовувати word2vec, кластирізацію Брауна (або кластери Брауна). Також, якщо брати N-грами, то можна використовувати для обчислення ваги слова NRC-10 словник з якого з якого заздалегідь виключити емоції та залишити емоційні класи.

1.2.2.1 NRC-10

Як сказано в [16], NRC словесно-емоційний асоціативний лексикон містить 14182 вручну анотованих англійські слова згідно з восьми емоціями: радість, смуток, злість, здивування, страх, огида, довіра, очікування; та двома емоційними класами: позитив та негатив. На рисунку 1.3 показано вид цього лексикону.

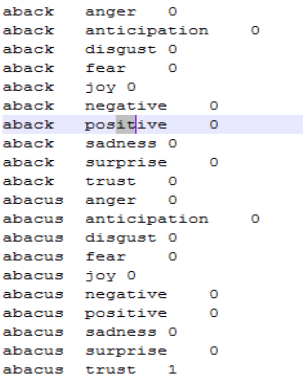

Рисунок 1.3 – NRC лексикон

1.2.2.2 Word2Vec

Word2vec [18, 19] – це інструмент (набір алгоритмів) для розрахунку векторних уявлень слів, який реалізує дві основні архітектури – Continuous Bag of Words (CBOW) і Skip-gram. На вхід подається корпус тексту, а на виході виходить набір векторів слів. Завданням методу CBOW є передбачення слова на підставі прилеглих слів. У skip-gram зворотна задача – пророкування набору прилеглих слів на підставі одного слова. На рисунку 1.4 показані ці архітектури.

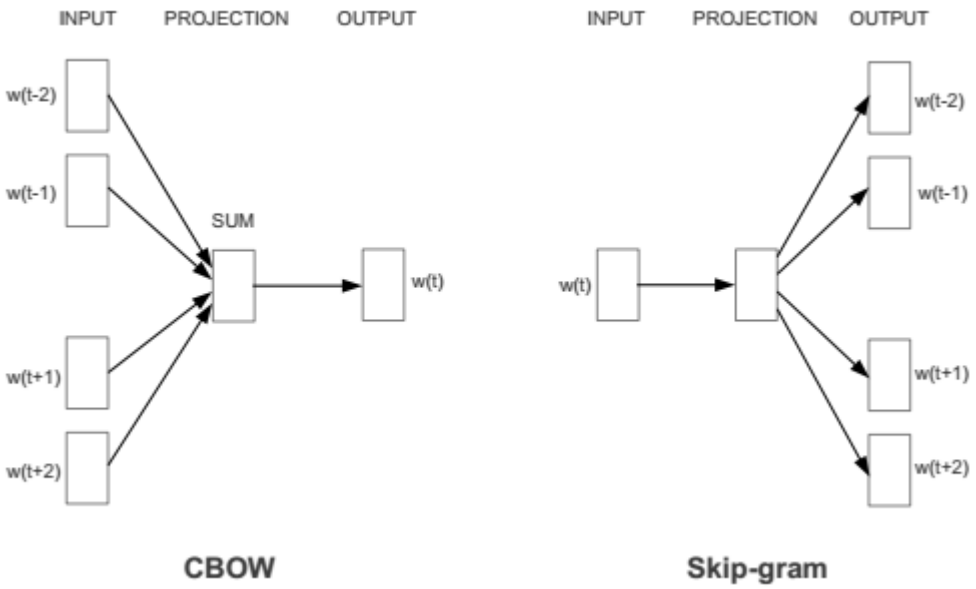

Рисунок 1.4 – CBOW та skip-gram

Обидва методи використовують в якості алгоритмів класифікації штучні нейронні мережі. Спочатку кожне слово в словнику – випадковий N-мірний вектор. Під час навчання алгоритм формує оптимальний вектор для кожного слова за допомогою методу CBOW або skip-gram.

Формально задача цього інструмента стоїть так: максимізація косинусної близькості між векторами слів (скалярний добуток векторів), які з'являються поруч один з одним, і мінімізація косинусної близькості між векторами слів, що не з'являються поруч один з одним. Поруч один з одним в даному випадку означає в близьких контекстах.

Наприклад у [16] використовується метод негативного семплінгу для тренування skip-gram моделі з цільового корпусу твітів. У цьому методі нейронна мережа з одним прихованим шаром навчається прогнозувати слова навколо центрального слова.

1.2.2.3 Brown clusters

Кожен твіт помічається згідно з кластерами слів Брауна та формується маломірний просторовий вектор у якому порахована частота кожного слова-кластера.

Кластеризація Брауна – складна ієрархічна агломераційна задача кастерізації на основі інформації про розподіл. Вона зазвичай застосовується до тексту, об’єднуючи слова в кластери, що семантично пов’язані між собою через застосування у схожих контекстах [20].

У [16] використовуються кластери Брауна з TweetNLP проекту.

На рисунку 1.5 показано один з варіантів кластерів Брауна. Перша колонка – путь кластера, друга – слово у цьому кластері, третя – частота, з якою зустрічається слово у твітах, які були застосовані для створення цих кластерів.

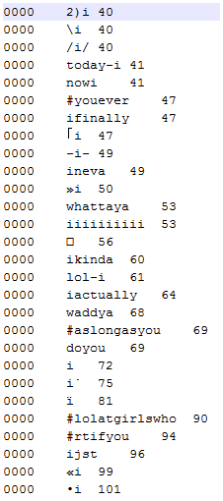

Рисунок 1.5 – Кластери Брауна

1.2.3 Інтелектуальний аналіз

Інтелектуальний аналіз тексту – напрям інтелектуального аналізу даних та штучного інтелекту, метою якого є отримання інформації з колекції документів. Інтелектуальний аналіз тексту використовує всі ті ж підходи до переробки інформації, що й Data Mining, однак між цими напрямками проявляється лише кінцевих методах, а також у тому, що Data Mining має справу з сховищами та базами даних, а не електронними бібліотеками та корпусами текстів.

Основними методами Data Mining є методи класифікації, прогнозування та моделювання, які основані на штучних нейронних мережах, деревах ухвалення рішень та ін. Найбільш впізнаваними є: метод опорних векторів, наївний Баєсів класифікатор, класифікатор Роше.

У загальному вигляді інтелектуальний аналіз можна відобразити у вигляді двох етапів (рис. 1.6):

— Тренувальний етап, під час якого завантажуються набори даних, що містять інформацію про контент та клас, якому він відповідає і виконується тренування та оцінка моделі.

— Етап прогнозування, під час якого тренована модель розгортується з новими наборами даних, які треба класифікувати.

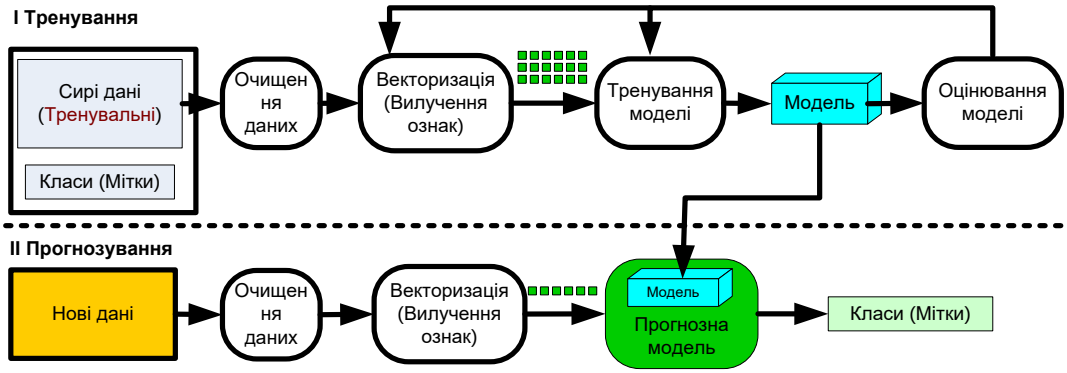

Рисунок 1.6 – Основні етапи роботи з текстовою аналітикою в середовищі контрольованого навчання

1.2.4 Класифікатори

Нижче описані класифікатори, які застосовувалися у [21] (вони є найбільш популярними), та логістична регресія (використовується у підрозділі 1.3.2).

1.2.4.1 Метод опорних векторів

Метод опорних векторів (SVM) працює за принципом розділення простору на підпростори, які відповідають класам. На рисунку 1.7 показано приклад такого розділення.

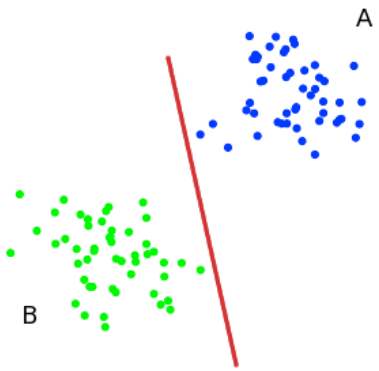

Рисунок 1.7 – Бінарна класифікація SVM

З точки зору точності класифікації вибирається пряма, відстань від якої до кожного класу максимально. Іншими словами, вибирається така пряма, яка розділяє класи найкращим чином. Ця пряма, а в загальному випадку – гіперплощина, називається

оптимальною розділяючою гіперплощиною, а вектори, що лежать найближче до цієї гіперплощини – опорними векторами.

1.2.4.2 Метод максимальної ентропії

Класифікатор максимальної ентропії є імовірнісним класифікатором, який належить до класу експоненціальних моделей. Цей метод базується на принципі максимальної ентропії і з усіх моделей, які відповідають навчальним даним та вибирає той, який має найбільшу ентропію. Він може використовуватися для вирішення великої кількості проблем класифікації тексту, таких як визначення мови, класифікації тем, аналізу настроїв тощо.

Загальний алгоритм побудови класифікатора, припускаючи, що ознаки використовується тільки рівня слів, можна описати так [24]:

— Для кожного ознаки/терма t та класу c ∈ C, визначити вектор f(c, d)

— За допомогою оптимізації призначити вагу кожному вектору, щоб максимально збільшити логічну вірогідність даних навчання.

— Імовірність класу c для документа d можна порахувати за формулою нижче.

$$Prob(c|d) = \frac{exp \sum_i^N w_i f_i(c,d)}{\sum_{c` \in C} exp \sum_i^N w_i f_i(c`,d)} \quad (1.6)$$

де Prob(c|d) – ймовірність класу c для документа d;

N – загальна кількість ваг для прогнозованого класу;

w_i – вага ознаки/терма для прогнозованого класу;

$f_i(c,d)$ – функція, яка повертає 1 або 0 за певних умов;

c`∈ C – кожен клас, який входить в множину класів C.

При прогнозі класу для нового документу потрібно знайти таке c`, щоб Prob(c|d) булла максимальною.

1.2.4.3 Наївний Байєсів класифікатор

Наївний Байєсів класифікатор, який базується на теоремі Байєса з наївним припущенням про те, що ознаки, які використовуються для класифікації – незалежні. Незважаючи на наївний дизайн та надто спрощені припущення, що використовує цей метод, Наївний Байєс добре працює у багатьох складних реальних проблемах [23].

Незважаючи на те, що цей класифікатор часто перевершує інші методи, він дуже ефективний, оскільки він менш обчислювальним (як у процесорі, так і в пам'яті), і вимагає невеликої кількості навчальних даних. Крім того, час навчання з Наївним Байєсом значно менший, ніж альтернативні методи.

У проблемі текстової класифікації можна використовувати слова (терми/токени) документа, щоб класифікувати його в відповідному класі. Загальний алгоритм навчання Наївного Байєса можна описати так [24]:

— Оцінити ймовірність Prob(c) кожного класу c ∈ C, розділивши кількість слів у документах з класом c на загальну кількість слів у корпусі.

— Оцінити розподіл ймовірностей Prob (t |c) для всіх слів t та класів c. Це можна зробити, розділяючи кількість t у документах з класом c загальною на кількість слів у класі c.

— Оцінити документ d для класу c можна за допомогою формули нижче.

$$score(d,c) = Prob(c) \cdot \prod_{i=1}^{n} Prob(t_i|c), \quad (1.7)$$

де t_i - слова (ознаки/терми) документа;

Prob(c) – ймовірність класу c;

Prob (t_i | c) – ймовірність для слова t та класу c.

Для того, щоб спрогнозувати найбільш ймовірну позначку класу, треба просто вибрати найбільший бал:

$$score = argmax\ Prob(c) \cdot \prod_{i=1}^{n} Prob(t_i|c) \quad (1.8)$$

1.2.4.4 Логістична регресія

Як сказано в [25] логістична регресія – відомий метод статистики, який використовується для прогнозування ймовірності результату, і особливо популярний для задач класифікації. Алгоритм передбачає ймовірність виникнення події за рахунок підбору даних для логістичної функції.

Цей метод так називається тому, що в ньому використовується логістична функція або сигмоїд (рис. 1.8):

$$f(z) = \frac{1}{1+e^{-z}} \quad (1.9)$$

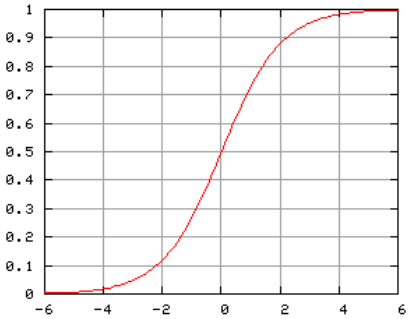

Рисунок 1.8 – Логістична функція

Як сказано в [26] логістична регресія використовує формулу дуже схожу на формулу лінійної регресії.

Вхідні значення (x) об'єднуються лінійно, використовуючи ваги коефіцієнта β для прогнозування вихідного значення (y). Ключовою відмінністю від лінійної регресії є те, що вихідні значення – бінарні (0 чи 1), а не числові. Нижче приведена приклад формули логістичної регресії.

$$y = \frac{e^{\beta_0+\beta_1 \cdot x}}{1+e^{\beta_0+\beta_1 \cdot x}}, \tag{1.10}$$

де y – прогнозоване значення;

β_0 – зміщення;

β_1 – коефіцієнт для вхідного значення x.

Формулу прогнозування можна представити наступним чином:

$$p(x) = \frac{e^{\beta_0+\beta_1 \cdot x}}{1+e^{\beta_0+\beta_1 \cdot x}} \tag{1.11}$$

чи

$$\ln\left(\frac{p(x)}{1-p(x)}\right) = \beta_0 + \beta_1 \cdot x \tag{1.12}$$

Коефіцієнти β оцінюються з тренувальних даних. Найкращі коефіцієнти приведуть до моделі, яка б передбачала значення, дуже близьке до 1 для класу за замовчуванням (або першого класу) та 0 до іншого класу.

У експерименті в підрозділі 1.1.2 використовувався модуль мультикласової логістичної регресії.

1.3 Аналіз та конкретизація задачі. Проведення початкового експерименту

Користувацький контент – різний інформаційно-значущий вміст носіїв інформації, що створюється користувачем.[1]

В основі будь-якого дослідження, додатку чи системи, що використовує дані з інтернету лежить Web Content Mining. Web Content Mining передбачає отримання корисної інформації з веб-контенту, такої як: текст, зображення, аудіо, відео та інше. Також, методи Web Content Mining дозволяють аналізувати цю інформацію [2].

Важливою інформацією, яку можна дістати та проаналізувати з користувацького контенту є емоції. Методи, які застосовуються для автоматизованого виявлення цих емоцій називаються методами аналізу тональності тексту чи аналізу думок (англ. sentiment analysis, opinion mining). Перші наукові роботи на цю тему датуються 2001-2002 роками [3, 4]

Для вирішення задачі аналізу тональності тексту використовують різні підходи:

— Підхід, заснований на словниках;

— Машинне навчання з вчителем (supervised learning);

— Машинне навчання без вчителя (unsupervised learning).

На даному етапі дослідження використовується машинне навчання з вчителем. Оскільки у цьому навчанні використовуються класифікатори, які навчаються на вже розмічених даних, та потім пророкують клас для нових даних, то визначення емоційного забарвлення можна звести до задачі класифікації.

Класифікація – задача, яка має безліч об’єктів, котрі розділені деяким чином на класи. Класифікувати об’єкт – вказати клас до якого належить даний об’єкт. Існує декілька типів класифікації [5]:

— Двокласова або бінарна класифікація. Найбільш простий випадок, який слугує для вирішення більш складних задач;

— Багатокласова класифікація. Число класів може доходити до кількох тисяч;

— Пересічні класи. Об’єкт може належати до кількох класів;

— Нечіткі класи. Визначається ступінь належності об’єкта до класу.

У користувацькому тексті з соціальних мереж можна виділити три класи: позитивний клас, негативний клас, нейтральний клас. Оскільки тут класів більше ніж два, то можна зробити висновок, що задача класифікації – багатокласова.

1.3.1 Розробка додатку для накопичування даних (твітів)

Був розроблений додаток, який за допомогою бібліотеки C# LinqToTwitter та StreamingAPI від Twitter отримує твіти, написані англійською мовою та записує у базу даних. Схема бази даних представлена на рис.1.9.

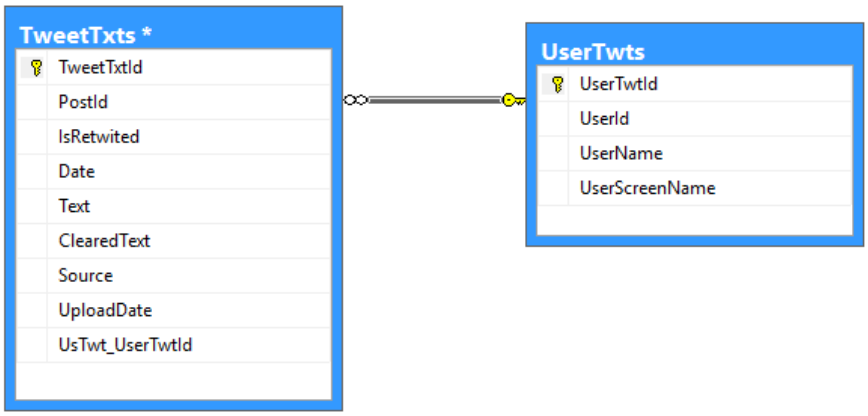

Рисунок 1.9 – Схема бази даних

База даних включає до себе дві таблиці: таблиця яка стосується безпосередньо твіта і таблиця користувачів, які написали/ретвітнули твіти.

За допомогою бібліотеки LinqToTwitter формується запит на створення потоку, який приймає лише англомовні твіти. Нижче наведено приклад такого запиту.

```
var st = await
    (from strm in context.Streaming
     where strm.Type == StreamingType.Sample && //(StreamingType.Filter поиск по ключевым словам)
     strm.Language == "en"
          select strm)
```

Сама функція отримання даних з Twitter показана у додатку Б (рядки 1-69).

Кожен запит, який формує потік, повертає JSON об’єкт. Цей об’єкт парситься бібліотекою Newtonsoft.Json.Linq (додаток А, рядок 26).

JObject obj = JObject.Parse(strm.Content);

З об’єкту ми отримуємо необхідну інформацію. Наприклад щоб отримати id поста треба написати obj["id"].

Цей додаток у середньому отримує за 5 хвилин ~3000 твітів або ~36000 за годину.

Якщо розглядати подальшу роботу у SMM потрібно створювати потік не лише за мовою, а і за ключовими словами. Streaming API не дає створювати потік за хештегами, що зменшує кількість твітів, що отримуються. Також якщо потрібно отримати твіти з

конкретним хештегом потрібно робити пошук серед вже отриманих твітів. Оскільки потік передає дані без критеріїв, то потрібна велика кількість твітв, щоб отримати потрібні.

Для пошуку за хештегами слід використовувати Search API. Приклад запиту показано нижче.

```
List<Status> searchResponse =
            await
            (from search in twitterCtx.Search
             where search.Type == SearchType.Search &&
                  search.Query == searchTerm && //ключевые слова и т.п.
                  search.Count == MaxSearchEntriesToReturn &&
                  search.SinceID == sinceID
             select search.Statuses)
                    .SingleOrDefaultAsync();
```

Перед подальшою роботою тексти твітів очищуються від непотрібної інформації: смайликів, символів @RT, посилань, хештегів, символів пунктуації. Це робиться за допомогою регулярних виразів. Наприклад вираз "[^\u0000-\u007F]+" дозволяє позбавитись від смайликів. Також слід позбавитись від стоп-слів – частих слів, які не несуть смислового навантаження, наприклад слова the, about, at та ін. Після попередньої обробки твіта виходить набір слів без інформаційного шуму.

Після можна провести процедури стемінга (виділення основи слова) та лематизації (привід слова до його нормальної форми). Якщо у тексті зустрічаються однакові слова, але з різними закінченнями, то за допомогою цих процедур їх можна привести до одного виду. Але це може не давати бажаних результатів. Наприклад слова «хочу» та «хотів» можуть мати різну тональність, оскільки в першому випадку може виражатись позитивні емоції та надія, а в другому – негативні, якщо автор виразив жаль.

1.3.2 Проведення експерименту

Був проведений експеримент у Microsoft Azure Machine Learning Studio (детальніше про MLS у підрозділі 1.1.3): була створена прогнозна та тренувальна модель модель (рис. 1.10 та рис 1.11).

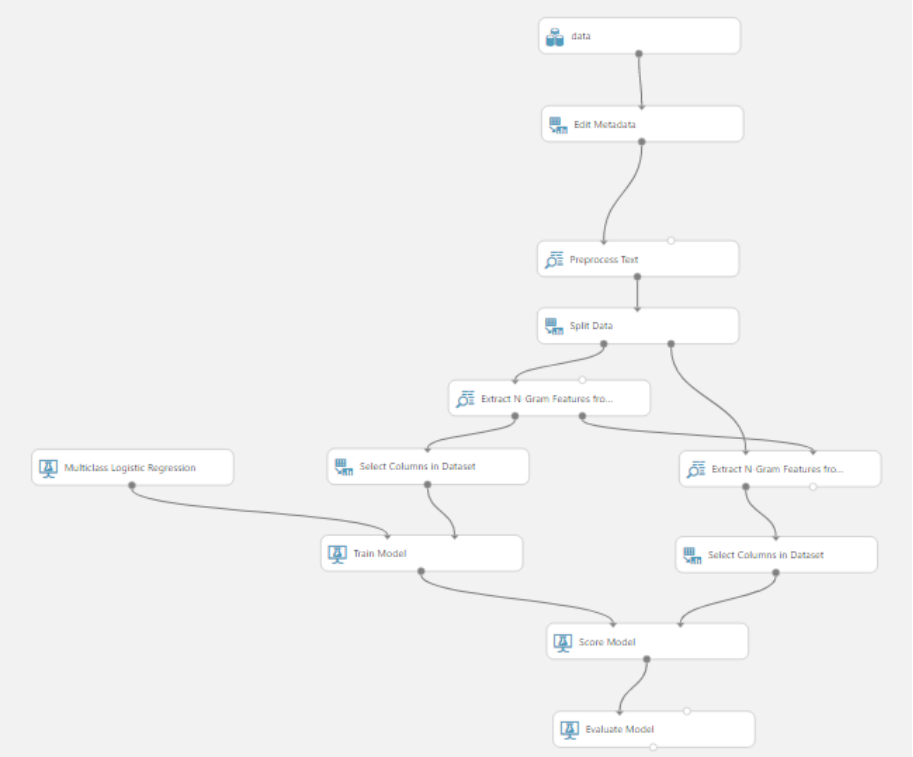

Рисунок 1.10 – Тренувальна модель

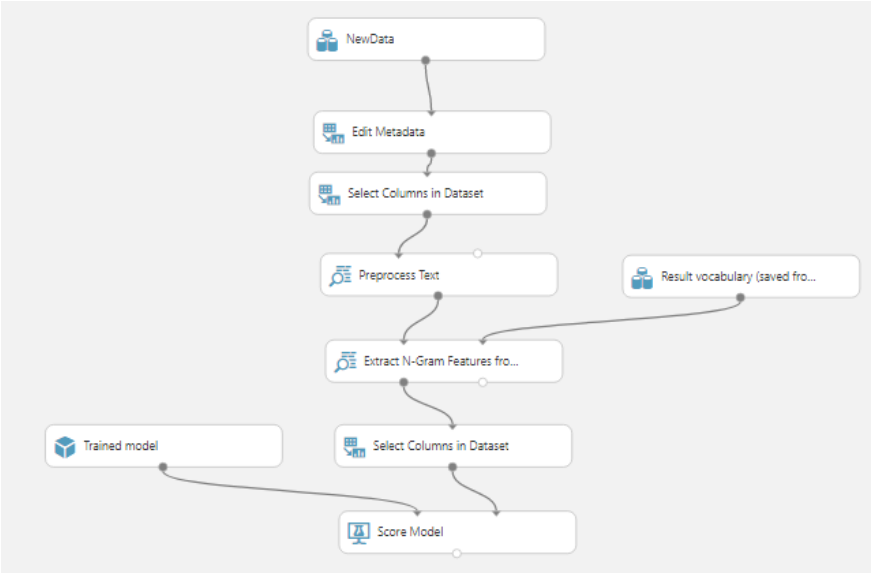

Рисунок 1.11 – Прогнозна модель

Data – дані (твіти) кількістю 4242, які були вручну розмічені залежно від емоцій в тексті, з яких 1939 (46%) помічені як нейтральні, 1328 (31%) – позитивні, 949 (22%) - негативні [6] (наглядно це показано на рисунку 1.12); NewData (прогнозна модель) – нерозмічені дані, які були отримані з підрозділу 1.1.1; Trained model та Result vocabulary – модель та словник n-gram (детальніше у підрозділі 1.2.) відповідно, які були отримані з експерименту з тренувальної моделі.

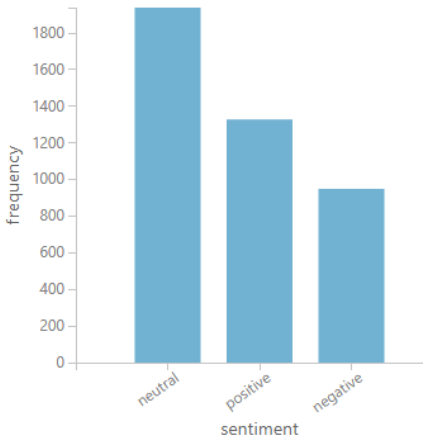

Рисунок 1.12 – Співвідношення класів з вхідних даних

Застосування тренувальної моделі на нерозмічених даних дало результат показаний на рисунку 1.13.

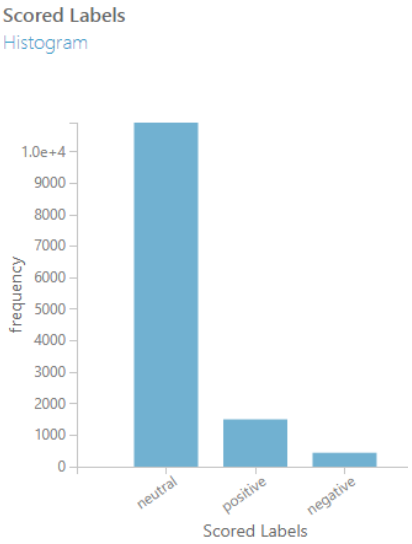

Рисунок 1.13 – Результат прогнозної моделі

10927 (85%) елементів визначило як нейтральні, 1502 (12%) – позитивні, 442 (3%) – негативні. Велика кількість елементів, що належать одному класу може говорити о незбалансованості даних [7].

1.3.3 Незбалансовані дані

Як сказано в [7], незбалансовані дані зазвичай стосуються проблеми класифікації, коли класи не представлені однаково. Наприклад, в бінарній класифікації використовуються 100 об’єктів, з яких 80 належать до одного класу, а 20 – до іншого.

Є задачі, коли класовий дисбаланс розглядається як не тільки загальний випадок, але й очікується. Наприклад, в набор даних, що характеризує шахрайські операції – незбалансований. Переважна більшість транзакцій відноситься до класу «не шахрайство», і дуже невелика – до «шахрайство».

Причиною отримання результату при якому отримується належність більшості об’єктів до одного класу є те, що модель вирішує завжди прогнозувати один клас [7]. Це дуже добре видно, якщо використовується алгоритм на простих заснований на простих правилах.

На рисунку 1.12 видно, що вхідні данні незбалансовані.

1.4 Постановка наукової задачі та обґрунтування методики досліджень

Результати проведеного аналізу моделей, методів, інструментальних засобів та після проведеного експерименту в підрозділі 1.3.2 були виявлені незбалансовані дані.

У відомій літературі [4, 16, 18, 21], задачі пов’язані з емоціональною класифікацією тексту в умовах незбалансованих даних не розглядалися.

В цьому контексті можна виділити наступні задачі:

— розробка метода, який би дозволив поліпшити результати емоційної класифікації;

— проведення експерименту для оцінки якості роботи метода.

Для проведення досліджень доцільно застосовувати методи, які б допомогли зменшити класовий дисбаланс.

1.5 Висновки до першого розділу

— Задача аналізу тональності тексту складна задача для вирішення якої застосовують різні методи у тому числі методи класифікації. Результати створення додатку для накопичування даних та подальшого експерименту показали, що існує проблема незбалансованих даних.

— Був проведений аналіз існуючих інструментів для взаємодії з Twitter та інструменту для машинного навчання (Microsoft Azure Machine Learning).

— Були проаналізовані базовий процес пошуку корисних знань в «сирих» даних та методи, що забезпечують цей процес.

— За результатами аналізів була сформована схема інформаційної технології (рис. 1.14).

— Сформульовані задачі магістерської роботи.

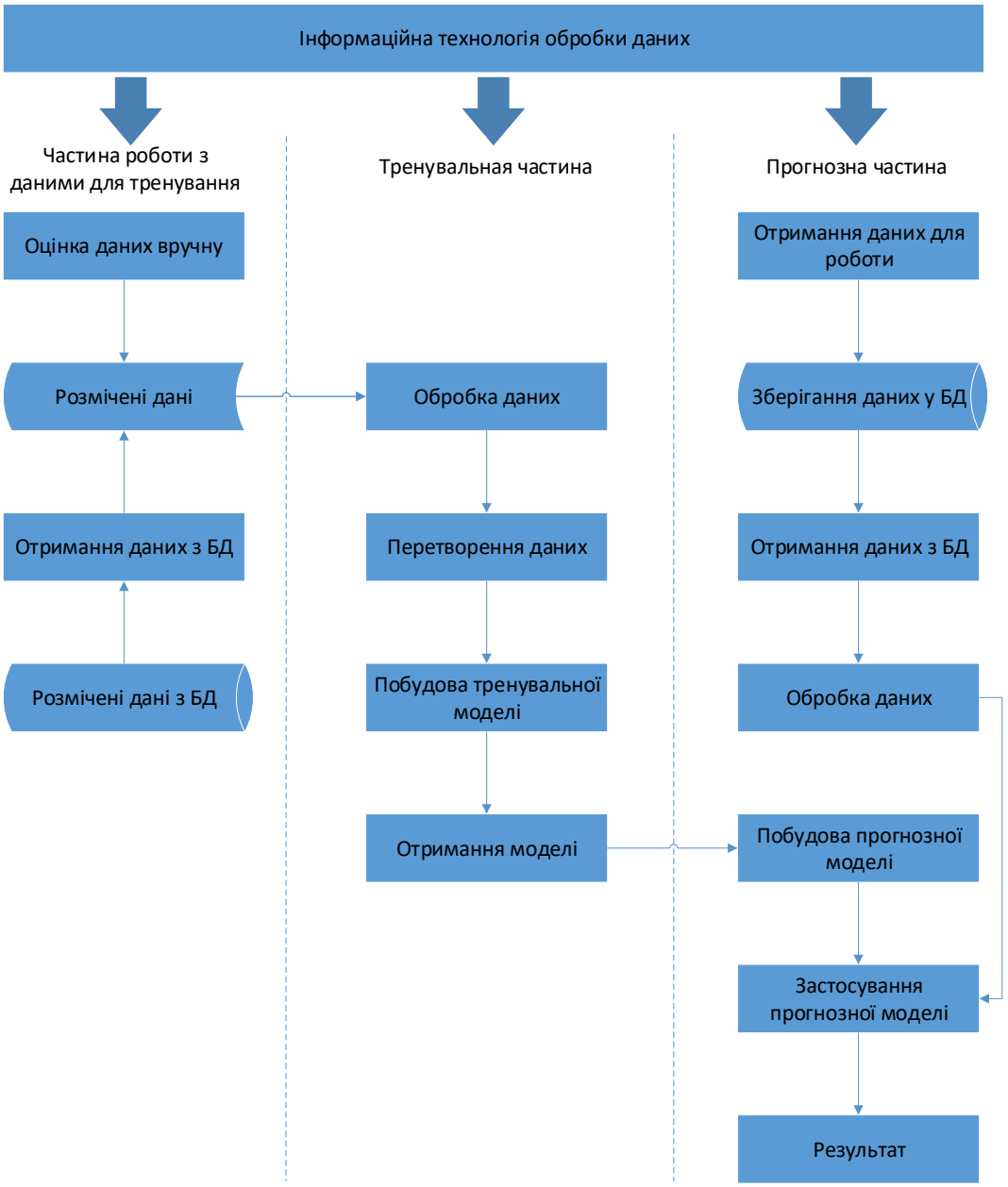

Рисунок 1.14 – Схема інформаційної технології обробки даних

2 МЕТОДОЛОГІЯ ПОБУДОВИ ТРЕНУВАЛЬНОЇ ТА ПРОГНОЗУЮЧОЇ МОДЕЛЕЙ КЛАСИФІКАТОРА

У розділі описана базова методологія побудови тренувальної та прогнозуючої моделей класифікатора та побудовані базові моделі у Microsoft Azure Machine Learning Studio.

2.1 Очищення та обробка даних

2.1.1 Регулярні вирази

Як було сказано в підрозділі 1.3.1 для того щоб ефективно застосовувати методи Data Mining, слід звернути увагу на питання попередньої обробки даних. Дані можуть містити пропуски, шуми, аномальні значення і т.д.

Для обробки даних можна використовувати регулярні вирази. Формально, ці вирази є алгебраїчним позначенням характеристик набору рядків. Також вони корисні для пошуку в текстах, коли є шаблон пошуку. Найпростіший вид регулярного виразу – проста послідовність символів. Приклад наведено в [27]: для того щоб знайти те чи інше слово застосовується команда /word/. Також команди регулярних виразів чутливі до реєстру: команда /word/ буде шукати слово, яке починається з маленької літери, а команда /Word/ - з великої.

З вхідних даних, тобто тексту з Twitter, на потрібно видалити наступну інформацію: посилання, смайлики, знак ‘@’, хештеги, вираз ‘RT’ та стоп-слова.

Стоп-слова – слова, які не несуть смислового навантаження [28]. До них можна віднести: суфікси, дієприкметники, прийменники, вигуки, цифри, нецензурна мова ті ін. Також іноді можна сюди віднести слова, які часто зустрічаються в інтернеті: інтернет, сайт, комп’ютер, прайс та ін.

Прикладом регулярних виразів для зазначених вище даних можуть слугувати:

— https?://[^]+ - для видалення посилань;

— [^\u0000-\u007F]+ - видалення смайликів;

— @+ - видалення знаку ‘@’;

— #+ - видалення хештега;

— \bRT\b – видалення виразу ‘RT’;

— /\b(as|at|he|the|was)\b/g – приклад регулярного виразу для видалення стоп-слів as, at, he, the, was.

Наразі більшість мов програмування, деякі текстові редактори та утиліти підтримують регулярні вирази [29]. Вид застосування виразів залежить від вибору мови програмування чи текстового редактора.

2.1.2 Лематизація та стемінг

Лематизація – приведення слова до його нормальної (словарної) форми. Наприклад: слова dinner та dinners мають однакову лему dinner [27]. Представлення слова його лемою важливо для веб-пошуку, тому що нам треба знаходити сторінки де не тільки лема, але й лематизоване слово.

Більшість методів для лематизації використовують морфологічний парсинг слова. Морфологія – вчення про будову слів та частини мови, основою одиницею якого є слово морфема в аспекті граматичної будови і творення [30]. Два головних класи морфем: основа та афікс. Наприклад: слово «dogs» має дві морфеми «dog» та «s», які морфологічний парсер повинен повернути.

Іноді застосовується стемінг – процес знаходження основи слова. Алгоритм стемінга, який найчастіше використовується – алгоритм Портера [31].

Як сказано в підрозділі 1.1.1 використання цих процесів може не давати бажаних результатів при використанні в аналізі тональності тексту. Наприклад слова «хочу» та «хотів» можуть мати різну тональність, оскільки в першому випадку може виражатись позитивні емоції та надія, а в другому – негативні, якщо автор виразив жаль. Тому, іноді ці процеси можна пропустити.

2.2 Перетворення даних

2.2.1 Векторизація

Наступним кроком в побудові базової тренувальної моделі є векторизація (вилучення ознак) з оброблених даних, отриманих в підрозділі 2.1, яка описана в підрозділі 1.3.2. Слід зазначити, що може скластися так, що ознака/терм не з’являються в жодному з документів і, щоб уникнути поділу на 0, до знаменника і чисельника формули 1.4 додають 1:

$$IDF(t, D) = log \frac{|D|+1}{|\{d_i \in D | t \in d_i\}|+1} \quad (2.1)$$

Також іноді встановлюють нижню границю 1, щоб уникнути ознаки/терми, що мають IDF 0:

$$IDF(t,D) = log\frac{|D|+1}{|\{d_i \in D|t \in d_i\}|+1} + 1 \quad (2.2)$$

Для формули 1.5 слід додати 1 до усіх елементів чисельника та знаменника:

$$w_{i,j} = F_{i,j} \cdot \log\left(\frac{(|Neg|+1)\cdot(Pos_t+1)}{(|Pos|+1)\cdot(Neg_t+1)}\right) + 1 \quad (2.3)$$

Весь цей процес називається згладжуванням (smoothing).

Розглянемо приклад базового розрахунку ваг слів, с застосуванням TF-IDF для двох документів: «The sky is blue» та «The sun in the sky is bright». Оскільки слова «the», «is», «in» - стоп-слова, то вони не беруть участь у розрахунку ваг. Залишається чотири слова: «sky», «blue», «sun» та «bright», - які формують словник N-грам. Підставивши значення для кожного документа та слова у формулу 1.3 можна отримати наступну матрицю:

$$TF_{matr} = \begin{bmatrix} 0{,}25 & 0{,}25 & 0 & 0 \\ 0{,}25 & 0 & 0{,}25 & 0{,}25 \end{bmatrix}, \quad (2.4)$$

де TF_{matr} – матриця значень TF для кожного слова у документах;

стовпці матриці – слова;

рядки – документи.

Підставивши значення у формулу 1.4, використовуючи натуральний логарифм, можна отримати наступний вектор:

$$IDF_{vec} = [0 \quad 0{,}6931 \quad 0{,}6931 \quad 0{,}6931], \quad (2.5)$$

де IDF_{vec} – вектор значень IDF для кожного слова;

кожен елемент вектору відповідає слову.

Тепер застосовується формула 1.2. Тобто треба помножити матрицю на матрицю. Для цього вектор idf_{vec} треба представити у вигляді діагональної матриці [32], яка показана у формулі 2.6.

$$IDF_{matr} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0{,}6931 & 0 & 0 \\ 0 & 0 & 0{,}6931 & 0 \\ 0 & 0 & 0 & 0{,}6931 \end{bmatrix} \quad (2.6)$$

Перемножуємо TF_{matr} та IDF_{matr}:

$$w_{matr} = TF_{matr} \cdot IDF_{matr} = \begin{bmatrix} 0 & 0{,}173275 & 0 & 0 \\ 0 & 0 & 0{,}173275 & 0{,}173275 \end{bmatrix}, \quad (2.7)$$

де w_{matr} – матриця ваг слів у веторно-просторовій моделі;

стовпці – слова;

рядки – документи.

Матрицю отриману в формулі 2.7 можна представити у вигляді таблиці 1.2. Нижче в таблиці 2.1 наведено результат готової векторно-просторової моделі для двох документів та чотирьох ознак/термів про які йшлося вище.

Таблиця 2.1 – Приклад готової векторно-просторової моделі

	Документ 1	Документ 2
sky	0	0
blue	0,173275	0
sun	0	0,173275
bright	0	0,173275

2.2.2 Нормалізація векторів

Нормалізація – приведення до одиничного розміру.

Значення TF з підрозділу 1.3.2 це в загальному випадку кількість входжень ознаки/терма у документ d. Якщо використовувати його у тому вигляді в якому воно описане, то це може призвести до проблеми спаму, коли один терм повторюється у документі з метою покращення його рейтингу [32]. Щоб подолати цю проблему частота документа зазвичай нормалізується. Наприклад, у формулі 1.3 нормалізація відбувається шляхом ділення числа входжень терма на загальну кількість слів у документі.

Нормалізація вектора – перетворення заданого вектора до вектора у тому ж напрямку, але з одиничною довжиною [32] (рис 2.1).

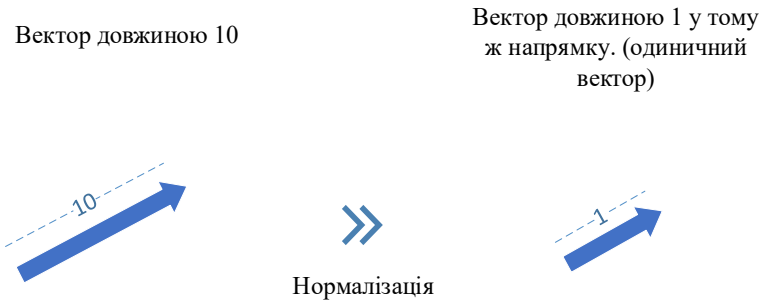

Рисунок 2.1 – Нормалізація вектора

Одиничний вектор розраховується за формулою:

$$\hat{v} = \frac{\vec{v}}{\|\vec{v}\|_p}, \tag{2.8}$$

де $\hat{v}$ – одиничний чи нормалізований вектор;

$\vec{v}$ – вектор, який потрібно нормалізувати;

$\|\vec{v}\|_p$ – норма вектора, який потрібно нормалізувати в L^p просторі.

Зазвичай довжина вектора розраховується за допомогою Евклідової норми. Евклідова норма – функція яка призначає строго позитивну довжину для всіх векторів у векторному просторі. Тому довжина розраховується за формулою:

$$\|\vec{v}\| = \sqrt{v_1^2 + \cdots + v_n^2} \tag{2.9}$$

Іноді біля норми є цифра p, $\|\vec{v}\|_p$, наприклад, тому формула 2.9 може бути представлена наступним чином:

$$\|\vec{v}\|_p = (|v_1|^p + \cdots + |v_n|^p)^{\frac{1}{p}} \tag{2.10}$$

Евклідову норму з p=2 називають L2-нормою. Вона використовується в інструментальних засобах в підрозділі 3.4.5.

2.2.3 Вибір ознак/термів

Також іноді застосовуються функції вибору ознак/термів, щоб отримувати лише ту інформацію, яка дійсно важлива. Ці функції присвоюють деякий показник ознаці/терму та потім вибирають лише найкращі ознаки. Прикладом такої функції може слугувати критерій хі-квадрат (chi-squared).

Хі-квадрат використовується в статистиці для перевірки незалежності двох подій. З вхідного набору даних отримуються спостережуваний O та очікуваний E показники. Хі-квадрат вимірює наскільки відрізняються ці показники [33]. В виборі ознак/термів події – присутність ознаки та присутність класу.

Якщо дві події залежні, появу ознаки можна використати для передбачення появи класу. Для навчання моделі треба вибрати ознаки/терми поява яких дуже залежить від виникнення класу.

Коли дві події незалежні спостережуваний показник близький до очікуваного, тому хі-квадрат дорівнює невеликому балу. Отже, високий бал хі-квадрат вказує на те, що гіпотеза про незалежність є хибною. Інакше кажучи, чим вище значення хі-квадрат, тим більше ймовірність, що ознака/терм для якої воно розраховується, співвіднесена з класом, отже, вона повинна бути вибрана для навчання моделі.

Розглянемо приклад розрахунку хі-квадрат для набору даних, які мають позитивний та негативний класи [33]. Екземпляр – документ.

Необхідні дані:

— A – кількість позитивних екземплярів, що містять ознаку t;

— B – кількість негативних екземплярів, що містять ознаку t;

— C – кількість позитивних екземплярів, що не містять ознаку t;

— D – кількість негативних екземплярів, що не містять ознаку t;

— M – кількість екземплярів, що містять ознаку t;

— P – кількість позитивних випадків;

— L – загальна кількість екземплярів.

Для обчислення необхідного значення для ознаки t можна збудувати таблицю:

Таблиця 2.1 – Данні для розрахунку хі-квадрат

	Позитивний клас	Негативний клас	Усього
Місять ознаку t	A	B	A+B=M
Не місять ознаку t	C	D	C+D=N-M
Усього	A+C=P	B+D=N-P	L

A, B, C, D – спостережувані показники.

Виходячи з гіпотези про те, що дві події є незалежними розрахувати, наприклад, очікуваний показник E_A можна за наступною формулою:

$$E_A = (A + C) \cdot \frac{A+B}{L} = P \cdot \frac{M}{L}, \tag{2.11}$$

Хі-квадрат розраховується за наступною формулою [34]:

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)}{E_i}, \tag{2.12}$$

де n – кількість показників;

O – спостережуваний показник;

E – очікуваний показник;

Підставивши значення у формулу 2.9 можна отримати наступне:

$$X^2 = \frac{(A-E_A)^2}{E_A} + \frac{(B-E_B)^2}{E_B} + \frac{(C-E_C)^2}{E_C} + \frac{(D-E_D)^2}{E_D} \tag{2.13}$$

Підставивши значення E_A, E_B, E_C, E_D та провівши деякі маніпуляцій можна отримати:

$$X^2 = \frac{L(AD-BC)^2}{(A+C)(B+D)(A+B)(C+D)} \tag{2.14}$$

З набору даних легко можна отримати значення A, M, P, N, тому формулу 2.11 можна представити наступним чином:

$$X^2 = \frac{L(AL-MP)^2}{PM(N-P)(L-M)}, \quad (2.15)$$

Для кожної ознаки розраховують за формулою 2.15 хі-квадрат та вибирають найкращі значення.

Також існують інші функції для вибору ознак/термів:

— Взаємна інформація (Mutual Information);

— Вибір ознак за частотою (Frequency-based feature selection);

— Функція на основі графів(Count Based);

— Різні кореляційні коефіцієнти;

— Та ін.

2.3 Тренування моделі

Як сказано в [27] логістична регресія для декількох класів називається мультиноміальною або багатокласовою, також іноді згадується в контексті, як максимальна ентропія (підрозділ 1.3.4.2). Формула 1.6 – загальна формула ймовірності класу c для документу x.

Розглянемо приклад бінарної (позитив чи негатив) класифікації документа, який має вигляд речення: «I do not like films of this director, but his last work is pretty good». Слова «not», «like», «pretty», «good» мають наступні ваги:

— not – 0,7 – негативне;

— like – 1,6 – позитивне;

— pretty – 1,4 – позитивне;

— good – 1,8 – позитивне.

Чим більша вага, тим позитивніше слово та навпаки: чим менше, тим негативніше.

Підставимо значення для кожного класу у формулу 1.6:

$$Prob(pos|x) = \frac{e^{1,6+1,4+1,8}}{e^{0,7}+e^{1,6+1,4+1,8}} = \frac{121,5104}{2,0137+121,5104} = 0,9836 \quad (2.16)$$

$$Prob(neg|x) = \frac{e^{0,7}}{e^{0,7}+e^{1,6+1,4+1,8}} = \frac{2,0137}{2,0137+121,5104} = 0,016 \quad (2.17)$$

З отриманих значень можна зробити висновок, що документ – позитивний тому, що значення 2.7 більше ніж 2.8.

Якщо мета – просто класифікація, то у формулі 1.6 можна проігнорувати знаменник та експоненту і просто вибрати клас з найвищою оцінкою:

$$Prob = argmax \sum_{i=1}^{N} w_i f_i(c,x) \quad (2.18)$$

Розрахунок для більш ніж двох класів такий саме за відмінністю того, що буде розраховуватися ймовірність ще для одного класу.

На відміну від, наприклад, наївного Байєсова класифікатора логістична регресія не передбачає незалежності ознак/термів, що дає можливість використовувати не тільки уніграми, а й біграми одночасно. Тому вона й використовується в експерименті в підрозділі 1.1.2 та наступних експериментах.

2.4 Побудова моделей

Якщо треба виповнити прогнозування для великої кількості даних, то краще використовувати готові засоби машинного навчання, такі як Microsoft Azure Machine Learning Studio.

2.4.1 Тренувальна модель

Першим кроком для побудови тренувальної моделі слугує створення нового експерименту в сервісі Microsoft Azure Machine Learning Studio. Для цього треба зайти на домашню сторінку Learning Studio та нажати New → Experiment → Blank Experiment. В результаті буте отримана чиста сторінка для роботи (рис. 2.2).

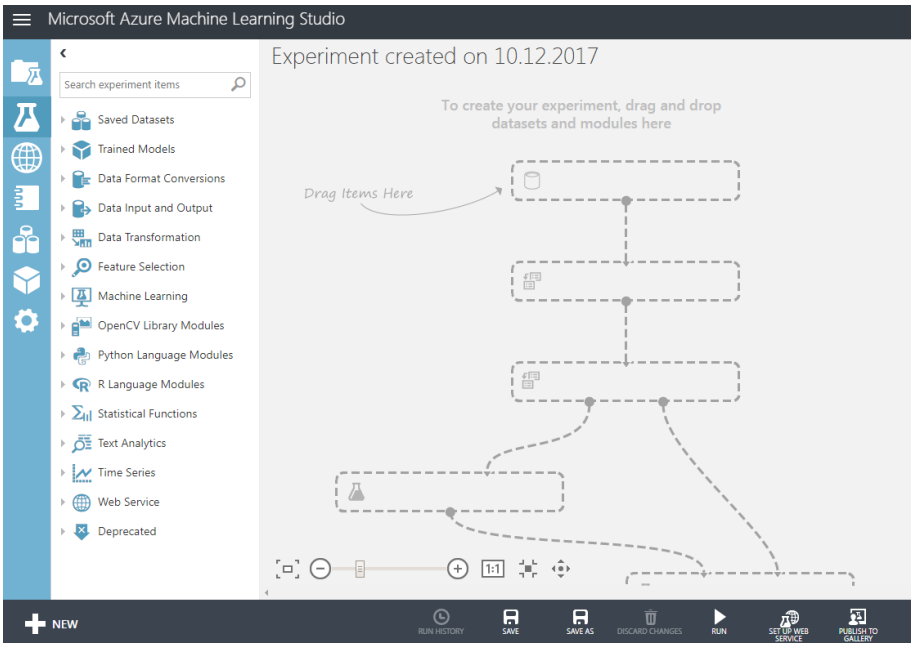

Рисунок 2.2 – Сторінка нового експерименту

Тепер треба завантажити тренувальні дані. У нашому випадку це дані з соціальних мереж та їх емоційна оцінка. Для цього треба нажати New → Dataset → From Local File, а потім в списку Select a type for the new dataset вибрати Generic CSV File with a header. Слід зазначити, що сервіс приймає дані формату csv, tsv, txt, arff, zip, RData, svmlight.

Після завантаження треба провести очищення та попередню обробку текстових даних. Якщо найменування колонок потребують перейменування, то застосовується модуль Edit Metadata (рис. 2.3).

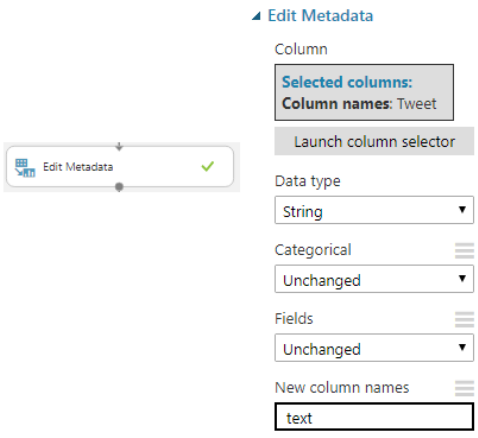

Рисунок 2.3 – Модуль Edit Metadata

Наступним модулем, який може застосовуватись може буди Group Categorical Values (рис 2.4), який групує значення. Наприклад, текстові дані мають оцінки у вигляді балів, а цей модуль може сформулювати дві категорії балів – низькі та високі.

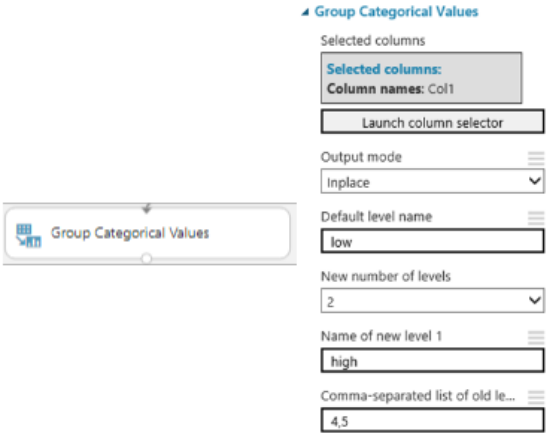

Рисунок 2.4 – Модуль Group Categorical Values

У експерименті в підрозділі 1.1.2 та наступних експериментах цей модуль не використовується, тому що дані мають емоційну оцінку у вигляді наступних значень: positive, negative та neutral.

Для очищення тексту використовується модуль Preprocess Text (рис 2.5). Очищення зменшує шум у наборі даних, допомагає знайти найважливіші функції та покращує точність кінцевої моделі.

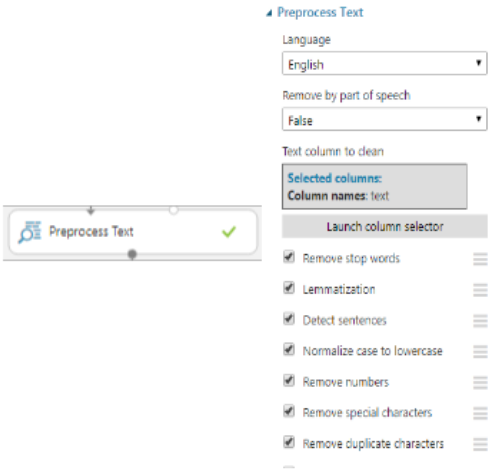

Рисунок 2.5 – Модуль Preprocess Text

Модуль дозволяє видалити стоп-слова - загальні слова, такі як "a" або "the" - і числа, спеціальні символи, дубльовані символи, адреси електронної пошти та URL-адреси.

У правому меню настройок модуля можна задати межі речень, які потім позначаються символом "|||" у попередньо обробленому тексті. Також в модулі можна використовувати регулярні вирази (підрозділ 2.1.1).

Після закінчення попередньої обробки обов’язково необхідно розбити дані на тренувальний і тестовий набори. Для цього використовується модуль Split Data (рис. 2.6).

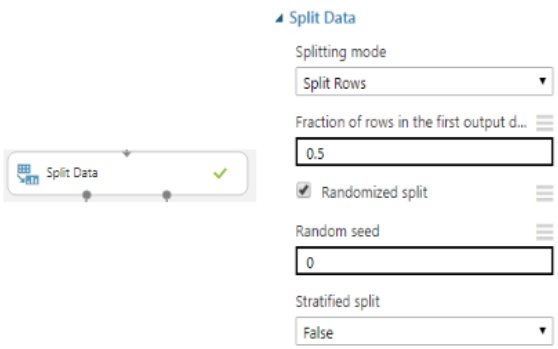

Рисунок 2.6 – Модуль Split Data

Далі треба перетворити дані у векторну модель. Для цього використовується модуль Extract N-Gram Features from Text (рис 2.7). Формально, блок приймає стовпчик зі словами, розділеними пробілами, і створює словник слів або N-грами слів, які містяться у наборі даних, якщо встановлений параметр Vocabulary mode у Create. Потім він підраховує, скільки разів кожне слово або N-грам з'являється у кожному записі, і створює векторні об'єкти з цих підрахунків.

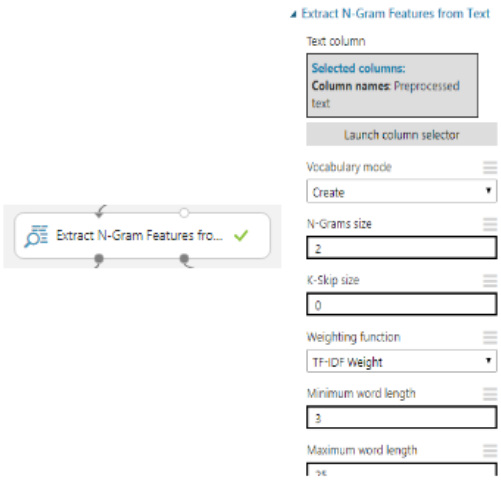

Рисунок 2.7 – Модуль Extract N-Gram Features from Text

Модуль дозволяє розраховувати ваги N-грам різними функціями. Однією з цих функцій є TF-IDF.

Такі текстові функції часто мають дуже високу розмірність. Наприклад, якщо збірка документів має 100 000 унікальних слів, то його простір буде мати 100 000 розмірностей або більше. Для зменшення розмірності встановлюються деякі параметри у поля Minimum n-gram document та Maximum n-gram document. Наприклад, для виключення N-грам, які з'являються у менш ніж 5 документах або у більш ніж 80% документів треба встановити Minimum n-gram – 5, а Maximum n-gram document – 0,8.

Модуль дозволяє використовувати функції вибору ознак/термів, щоб зменшити їх кількість та використовувати лише найважливіші. В ньому представлені наступні функції:

— Коефіцієнт кореляції Пірсона (Perason Correlation);
— Взаємна інформація (Mutual Information);
— Коефіцієнт кореляції рангу Кендала (Kendall Correlation);
— Коефіцієнт кореляції рангу Спірмена (Spearman Correlation);
— Хі-квадрат (Chi Squared);
— Оцінка Фішера (Fisher Score);
— Функція на основі графів(Count Based);

Як альтернативний підхід до використання функцій Extract N-gram Features from Text, можна використовувати модуль Feature Hashing. Слід звернути увагу, що останній не має вбудованих можливостей вибору функцій, або TF-IDF зважування.

Як сказано вище, модуль Extract N-gram Features from Text генерує словник N-грам (рис. 2.8), який використовується в перевірці тренувальної моделі та в прогнозуючій моделі.

Id	Ngram	DF	IDF
1	count	5	2.627571
2	steve	5	2.627571
3	water	5	2.627571
4	roll	5	2.627571

Рисунок 2.8 – Словник згенерований модулем Extract N-gram Features from Text

Також Extract N-Gram Features from Text може допомогти розширити існуючий словник якщо встановити параметр Vocabulary mode у Merge, або оновити його, встановивши параметр Vocabulary mode у Update.

Наступним кроком є тренування моделі класифікації. На цьому кроці текст вже перетворений на стовпці з числами. Однак, набір даних все ще містить рядки стовпців з попередніх етапів, тому слід використати модуль Select Columns in Dataset, щоб виключити їх.

Для прогнозування оцінок використовується модуль логістичної регресії. В сервісі Microsoft Azure Machine Learning Studio він представлений двома модулями: Two-Class Logistic Regression та Multiclass Logistic Regression, - які застосовуються для задач бінарної та багатокласової класифікації відповідно. У експерименті в підрозділі 1.1.2 та наступних експериментах використовується Multiclass Logistic Regression (рис 2.9).

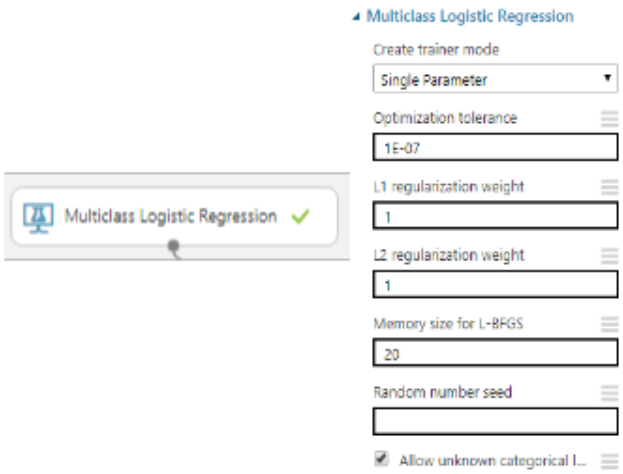

Рисунок 2.9 – Модуль Multiclass Logistic Regression

Для перевірки моделі до другого виходу модуля Split Data треба під’єднати новий модуль Extract N-gram Features from Text до другого входу якого під’єднати другий вихід модуля модуль Extract N-gram Features from Text з навчальної гілки та встановити режим словника ReadOnly (рис. 2.10).

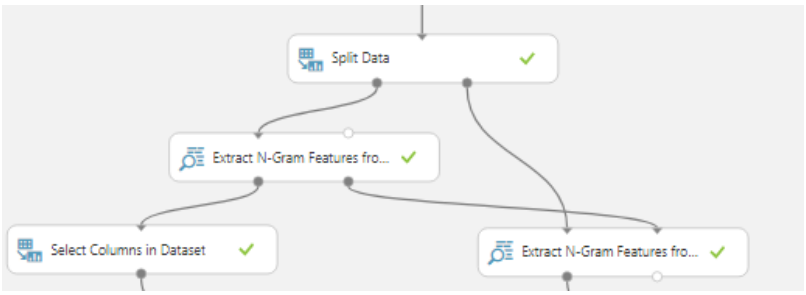

Рисунок 2.10 – Перевірка моделі

Також необхідно вимикнути фільтрацію N-грам за частотою, встановивши Minimum n-gram 1 та Maximum n-gram document 1, і вимкнути User filter-based feature selection вибравши зі списку False.

Після того, як текстовий стовпець даних тесту перетворено на стовпці з числовими значеннями, виключаються стовпці з попередніх етапів, використовуючи модуль Select Columns in Dataset.

Щоб зробити прогнозування та оцінити точність використовується модуль Score Model.

Також в кінці можна добавити модуль Evaluate Model. Цей модуль дозволяє виміряти точність навченої моделі, використовуючи дані з модуля Score Model. Повертає стандартні показники оцінки такі, як точність, чутливість, матриця неточностей, які розглянуті більш детально в підрозділі 3.1.

Підсумувавши все вище зазначене можна отримати тренувальну модель з підрозділу 1.1.2, яку показано на рисунку 1.2.

Така модель майже готова до використання. Якщо її розгорнути як веб-сервіс, то вона зможе отримувати на вході текстовий рядок у вільному вигляді, і повертати на виході прогноз. Для прогнозування вона буде використовувати словник N-грам для перетворення тексту, а також навчену логістичну регресійну модель, щоб зробити прогноз.

2.4.2 Прогнозна модель

В якості основи для створення прогнозної моделі можна використати попередньо створену модель (підрозділ 2.4.1 та рис. 1.2), але виконати деякі перетворення.

Отже, щоб зробити прогнозний експеримент, спочатку треба зберегти словник N-грам як набір даних (N-Gram Vocabulary) і навчену модель логістичної регресії з навчальної гілки експерименту (Train Model).

Далі треба створити пустий експеримент, як в підрозділі 2.4.1 та завантажити дані для яких треба зробити прогноз. Потім додати необхідні модулі (Edit Metadata, Select Columns in Dataset, Preprocess Text та Extract N-Gram Features from Text). В останньому модулі треба вибрати такі ж налаштування, як в підрозділі 2.4.1 з гілки перевірки моделі та до другого входу слід під’єднати збережений словник.

Останнім кроком є під’єднання до першого входу модуля Score Model збереженої моделі логістичної регресії з навчальної гілки тренувального експерименту.

Підсумувавши все вище зазначене можна отримати прогнозну модель з підрозділу 1.1.2, яку показано на рисунку 1.3.

2.5 Висновки до другого розділу

В розділі «Методологія побудови тренувальної та прогнозуючої моделей класифікатора» були розглянуті основні кроки при проектуванні моделей та розглянутий приклад їх побудови в сервісі Microsoft Azure Machine Learning Studio.

3 МЕТОДИ ОЦІНКИ МОДЕЛЕЙ ПРОГНОЗУВАННЯ ТА МЕТОДИ БОРОТЬБИ З НЕЗБАЛАНСОВАНИМИ ДАНИМИ

У розділі проаналізовані методи оцінки моделей прогнозування та методи боротьби з незбалансованими даними, розроблені алгоритми для реалізація міри дельта TF-IDF та описана сама реалізація міри дельта TF-IDF.

3.1 Засоби оцінки ефективності моделей класифікації

Основою перевірки моделі є тестова вибірка в якій визначена відповідність між документами і їх класами. Для оцінки роботи того чи іншого класифікатора достатньо застосувати класифікатор, на цій вибірці та співвіднести його рішення з правильними рішеннями та використати чисельні метрики, які описані нижче.

3.1.1 Матриця неточностей

Матриця неточностей надає результат хибно позитивних, хибно негативних, істинно позитивних та істинно негативних результатів. Матриця неточностей показує кількість правильних і неправильних прогнозів, зроблених моделлю порівняно з фактичними класами. Ефективність моделей класифікації зазвичай оцінюється з використанням даних у матриці розмірністю N×N, де N - кількість класів. Приклад матриці показано в таблиці 3.1 [35].

Таблиця 3.1 – Приклад матриці неточностей

N=165	Прогнозований клас: позитивний	Прогнозований клас: негативний
Фактичний клас: позитивний	50	10
Фактичний клас: негативний	5	100

Ця матриця надає нам такі дані:

— Присутні два класи «позитивний» та «негативний» (при прогнозуванні настрою перший клас означає, що документ позитивний, а другий – негативний);

— Класифікатор зробив 165 прогнозувань;

— Зі 165 прогнозувань класифікатор спрогнозував 50 позитивних класів, та 100 негативних;

— Фактично 105 документів негативні, а 60 позитивні.

Основні терміни:

— true positives (TP): Випадки у яких спрогнозовано позитивний клас.

— true negatives (TN): Випадки у яких спрогнозовано негативний клас.

— false positives (FP): Випадки у яких спрогнозовано позитивний клас, але фактично клас негативний.

— false negatives (FN): Випадки у яких спрогнозовано негативний клас, але фактично клас позитивний.

Тепер таблицю 3.1 можна представити наступним чином:

Таблиця 3.2 – «Модернізована» матриця неточностей

N=165	Прогнозований клас: позитивний	Прогнозований клас: негативний	
Фактичний клас: позитивний	TP=50	FN=10	60
Фактичний клас: негативний	FP=5	TN=100	105
	55	110	

3.1.2 Accuracy, precision, recall

Точність (accuracy) – частка документів за якими класифікатор прийняв правильне рішення. Загальна точність (overall accuracy) на прикладі таблиці 3.2 обчислюється за наступною формулою [36]:

$$Accuracy = \frac{P}{N} = \frac{TP+TN}{total}, \quad (3.1)$$

де P – кількість документів за якими класифікатор прийняв правильне рішення;

N – розмір тренувальної вибірки;

total – загальна кількість прогнозів.

У загальному вигляді точність можна обрахувати за наступною формулою:

$$Accuracy = \frac{\sum d_{correct}}{N}, \quad (3.2)$$

де $d_{correct}$ – кількість документів у яких клас був визначений правильно;

N- загальна кількість документів.

Також існує середня точність (average accuracy), яка застосовується при мультикласовій класифікації. Вона обчислюється за формулою:

$$Accuracy_{avg} = \frac{\sum_{i=1}^{k} Accuracy}{k}, \quad (3.3)$$

де чисельник – сума точностей для кожного класу;

k – загальна кількість класів.

Чутливість (recall) – відсоток (частка) позитивних випадків, які були правильно класифіковані як позитивні. Обраховується за формулою на прикладі таблиці 3.1 [37]:

$$Recall = \frac{TP}{TP+FN} \quad (3.4)$$

Precision – показує наскільки часто класифікатор прогнозує позитивні випадки. Обраховується за формулою на прикладі таблиці 3.1 [37]:

$$Precision = \frac{TP}{TP+FP} \quad (3.5)$$

У випадку мультикласової класифікації використовуються методи micro-average та macro-averaged по відношенню до recall та precision [38].

В методі micro-average підсумовуються індивідуальні правильні позитивні (TP), помилкові позитивні (FP) та помилково негативні (FN) випадки для різних наборів даних. Recall та Precision обраховуються за формулами:

$$Recall_{mic-avg} = \frac{TP_1+\cdots+TP_k}{TP_1+\cdots+TP_k+FN_1+\cdots+FN_k}, \quad (3.6)$$

$$Precision_{mic-avg} = \frac{TP_1+\cdots+TP_k}{TP_1+\cdots+TP_k+FP_1+\cdots+FP_k}, \quad (3.7)$$

де TP_1, TP_k – правильні позитивні випадки для кожного набору даних відповідно;

FN_1, FN_k – помилково негативні випадки для кожного набору даних відповідно;

FP_1, FP_k – помилкові позитивні випадку для кожного набору даних відповідно;

k – кількість класів.

Метод macro-averaged також застосовується для різних наборів даних. Для кожного набору окремо обраховуються, підсумовуються метрики та діляться на кількість класів. Представити це можна у вигляді наступних формул на прикладі двох наборів даних:

$$Recall_{mac-avg} = \frac{Recall_1+\cdots+Recall_k}{k}, \quad (3.8)$$

$$Precision_{mac-avg} = \frac{Precision_1+\cdots+Precision_k}{k} \quad (3.9)$$

3.2 Методи боротьби з незбалансованими даними засновані на кількості елементів в класах

3.2.1 Збільшення кількості початкових даних для тренування моделі

В [7] запропоновано найпростіший спосіб боротьби з небалансованими даними: зібрати більше даних. Великий набір даних може зберігати в собі більш збалансовані класи.

У першому експерименті було використано тільки розмічені дані кількістю 4242 з Twitter, але для тренування моделі можна використовувати й інші невеликі розмічені тексти, коментарі з YouTube чи інших соцмереж.

Були взяті дані з таких сайтів: YouTube, Twitter, Runner World forum, Digg, BBC, MySpace, - загальною кількістю 11794 текстів. З усіх даних 4476 (38%) помічені як позитивні, 4003 (34%) – нейтральні, 3289 (28%) – негативні.

Був проведений подібний експеримент як у підрозділі 1.1.3. Порівняння співвідношення класів з експерименту в підрозділі 1.1.2 показано на рисунку 3.1.

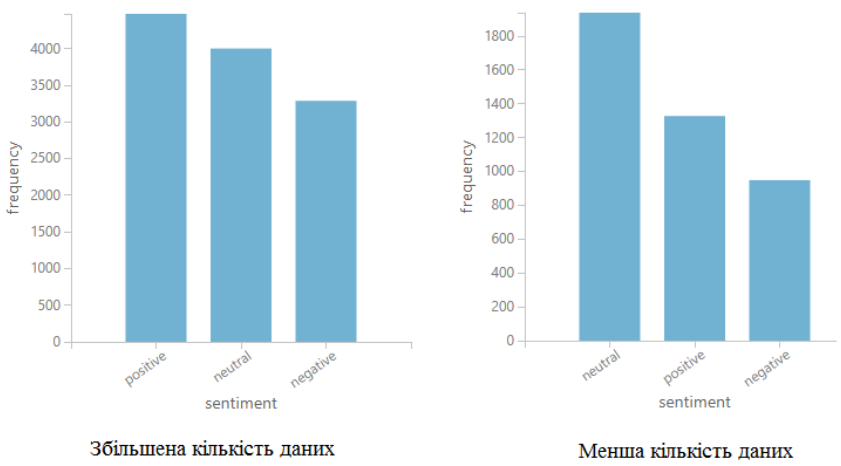

Рисунок 3.1 – Порівняння співвідношення класів

З рисунку 3.1 видно, що більша кількість даних має більш збалансовані класи.

Модуль Evaluate Model в Microsoft Azure Machine Learning Studio для оцінки моделі використовує метрики описані вище, а саме: overall accuracy, average accuracy, micro-averaged precision, macro-averaged precision, micro-averaged recall, macro-averaged recall та матрицю неточностей.

Оцінкою моделі можуть слугувати рисунок 3.2 на якому зображена матриця неточностей та таблиця 3.3.

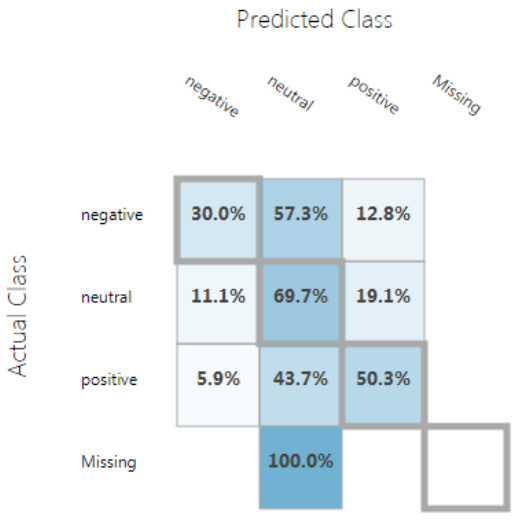

Рисунок 3.2 – Матиця неточностей для більшої кількості вхідних даних

Таблиця 3.3 – Метрики для оцінки моделі з експерименту з більшою кількістю вхідних даних

Overall accuracy	Average accuracy	Micro-averaged precision	Macro-averaged precision	Micro-averaged recall	Macro-averaged recall
0,512464	0,756232	0,512464	NaN	0,512464	0,37502

З матриці неточностей видно, що класифікатор правильно спрогнозував 30% негативних документів, 69,7% нейтральних та 50,3% позитивних.

Застосування нової тренувальної моделі на нерозмічених даних дало результат показаний на рисунку 3.3.

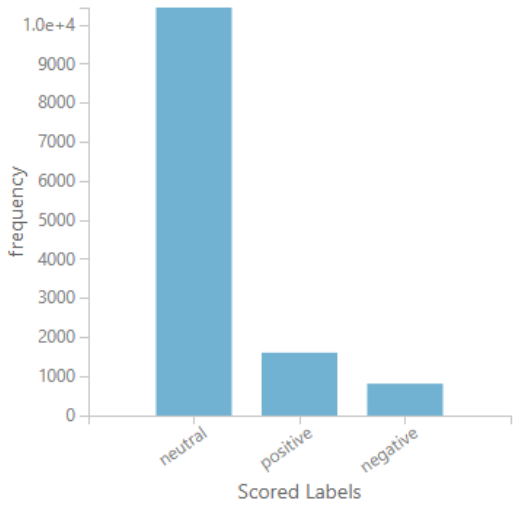

Рисунок 3.3 - Результат прогнозної моделі для більшої кількості вхідних даних

10448 (81%) елементів визначило як нейтральні, 1608 (12%) – позитивні, 815 (6.3%) – негативні. Отже, більш великий набір даних, у порівнянні з даними з експерименту в підрозділі 1.1.2, дозволив зменшити кількість елементів, що відносяться до найбільшого класу на 4% та трохи збільшити кількість елементів в і інших класах.

3.2.2 Використання збалансованих початкових даних для тренування моделі

Також для навчання моделі можна використовувати вже збалансовані дані: однакова кількість елементів належать до різних класів. Для вибору однакової кількості

текстів були застосовані стандартні інструменти Microsoft Excel. Було отримано 9828 текстів (по 3276 у кожному класі).

Оцінкою моделі можуть слугувати рисунок 3.4 на якому зображена матриця неточностей та таблиця 3.4.

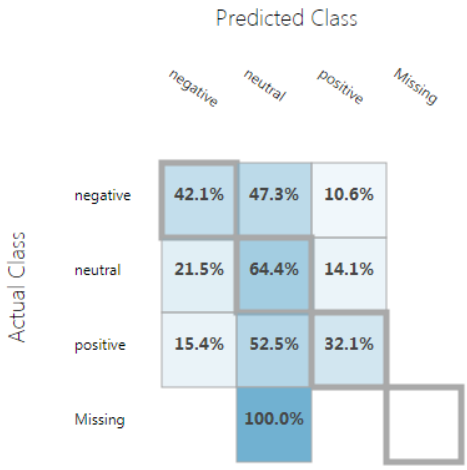

Рисунок 3.4 – Матриця неточностей для збалансованих вхідних даних

Таблиця 3.4 – Метрики для оцінки моделі з експерименту зі збалансованими вхідними даними

Overall accuracy	Average accuracy	Micro-averaged precision	Macro-averaged precision	Micro-averaged recall	Macro-averaged recall
0,459602	0,729801	0,459602	NaN	0,459602	0,346439

З матриці неточностей видно, що класифікатор правильно спрогнозував 42,1% негативних документів, 64,4% нейтральних та 32,1% позитивних.

Застосування нової тренувальної моделі на нерозмічених даних дало результат показаний на рисунку 3.5.

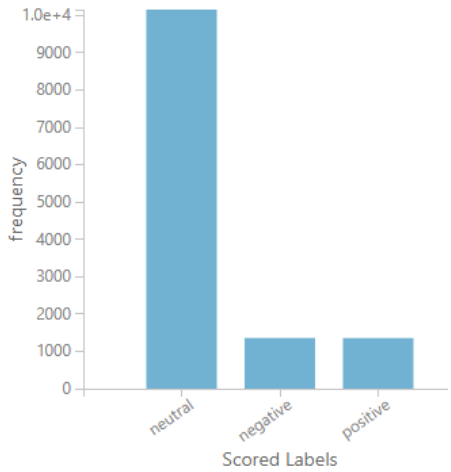

Рисунок 3.5 - Результат прогнозної моделі при використанні збалансованих даних

10158 (79%) елементів визначило як нейтральні, 1358 (11%) – позитивні, 1355 (10.9%) – негативні. Початкові збалансовані дані, , у порівнянні з даними з експерименту в підрозділі 1.1.2, дозволили зменшити кількість елементів, що відносяться до найбільшого класу на 6% та вирівняти показники позитивного та негативного класів (11%).

3.3 Застосування міри дельта TF-IDF для обчислення ваг ознак/термів

Ще одним варіантом роботи з незбалансованими даними є застосування на навчальному етапі методу обчислення ваг термів дельта TF-IDF замість просто TF-IDF. Ідея цього метода полягає у тому, щоб дати більшу вагу словам, які мають не нейтральну тональність.

Цей метод описано у [22]. У цій статті автори зазначають, що, наприклад, у моделі bag of words кожне слово чи пара слів асоціюється з деякою величиною. Зазвичай величина – число входжень цього слова у документ. Те саме можна сказати і про векторно-просторову модель (підрозділ 1.3.2), з поправкою на те, що величина вимірюється такими методами як булева вага, tf та tf-idf.

Автори статті пропонують обчислювати вагу (величина, що асоціюється зі словом/ознакою/термом) ознаки/терма шляхом обчислення результатів TF-IDF цієї ознаки/терма у позитивних та негативних тренувальних даних.

Для обчислення IDF можна застосувати наступну формулу:

$$IDF = \log\frac{|Pos|}{Pos_i} - \log\frac{|Neg|}{Neg_i} \tag{3.10}$$

Обчислити ваги ознак/термів окрім формули 1.5 можна за формулою:

$$w_{i,j} = F_{i,j} * \log\left(\frac{|Pos|}{Pos_i}\right) - F_{i,j} * \log\left(\frac{|Neg|}{Neg_i}\right) =$$
$$= F_{i,j} * \log\left(\frac{|Pos|Neg_i}{Pos_i|Neg|}\right) = F_{i,j} * \log\left(\frac{Neg_i}{Pos_i}\right) \quad (3.11)$$

Ознаки, що були більше помічені у негативних тренувальних даних, чим у позитивних мають позитивне значення та навпаки, ознаки помічені у позитивних – негативне.

3.3.1 Реалізація дельта TF-IDF

Дельта TF-IDF представлена бібліотекою sklearn-deltatfidf для мови програмування Python. Вона заснована на [22] та найчастіше використовується для класифікації настроїв. Успадковує бібліотеку sklearn-tfidf, яка в свою чергу є частиною бібліотеки машинного навчання scikit-learn [39] для мови Python.

Успадкування – механізм утворення нових класів на основі існуючих в об'єктно-орієнтовному програмуванні при якому властивості батьківського класу переходять до нащадка.

Згідно з прикладом застосування у [40] для обчислення ваг ознак вихідні дані (документи) та оцінки до цих даних представлені у вигляді списків, де позиція оцінки відповідає позиції документа. Оцінки дорівнюють наступним значенням:

— 1 – документ позитивний;

— -1 – документ негативний.

3.3.2 Приклад розрахунку дельта TF-IDF

Візьмемо, наприклад, слово «comment», яке зустрічається у 36 позитивних документах з 2238 та 22 негативних документах з 1665 розширених вхідних даних з підрозділу 3.2. Для обчислення IDF підставимо необхідні значення у формулу 3.10.

$$IDF = \log\frac{2238}{36} - \log\frac{1665}{22} = \log 62{,}166 - \log 75{,}6818 = -0{,}1967 \quad (3.12)$$

Порахуємо значення дельта TF-IDF для слова «comment», наприклад для документа №66 з розширених та оброблених вихідних даних: «degree seperation you fortunate close artistic treasure thank share your comment», - підставивши значення у формулу 3.12.

$$w_{i,j} = 0{,}0909 * \log 62{,}166 - 0{,}0909 * \log 75{,}6818 = -0{,}0179 \quad (3.13)$$

Перш ніж проводити наступний експеримент були розроблені алгоритми за якими виконувалася побудова додаткового ПЗ.

3.3.3 Розробка алгоритмів

Для створення даних для тренування у Microsoft Azure Machine Learning Studio треба виконати дії показані на рисунку 3.6. Блок схема алгоритму наведена в додатку А, рис. А1-А2.

Рисунок 3.6 – Послідовність дій для створення даних для тренування моделі

Також якщо кількість даних дуже велика, то можна розбити початкові дані на декілька блоків та розбити дії, що показані на рисунку на гілки та в кінці об'єднати в один файл. На рисунку 3.7 показано приклад при розбитті даних на дві гілки.

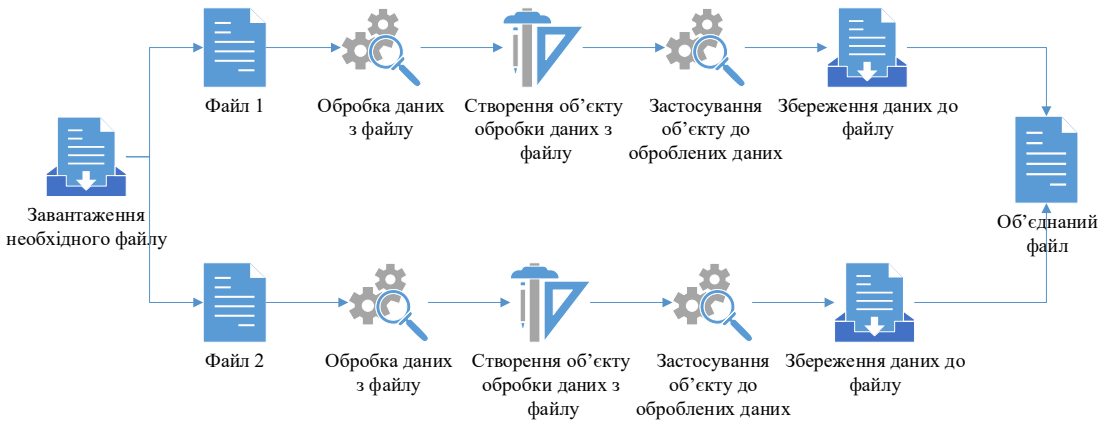

Рисунок 3.7 - Послідовність дій для створення даних для тренування та оцінки моделі при великій кількості даних

Крок «застосування об’єкту до оброблених даних» повертає не тільки оброблені дані, але й словник, який включає до себе ознаку/терм, DF (document frequency) – кількість документів у яких присутня ознака, та IDF. Такий словник можна застосовувати в Microsoft Azure Machine Learning Studio, оскільки модуль Extract N-Gram Features from Text приймає тільки значення зазначені вище.

Нижче на рисунку 3.8 показана послідовність дій з використанням словника для створення тестових даних для оцінки моделі. Блок-схема алгоритму наведена в додатку А, рис. А4.

Рисунок 3.8 - Послідовність дій для створення даних для оцінки моделі

3.3.4 ПЗ, мова програмування та бібліотеки для використання дельта TF-IDF

Для використання дельта TF-IDF було використано додаткові інструментальні засоби: мова програмування Python; Pycharm Community Edition – середовище розробки для мови програмування Python; scikit-learn – бібліотека машинного навчання для мови Python; sklearn-deltatfidf – бібліотека для використання дельта TF-IDF; openpyxl – бібліотека для взаємодії з файлами Microsoft Excel.

3.3.4.1 Python

Python – інтерпретована об’єктно-орієнтовна мова програмування високого рівня зі строгою динамічною типізацією [41], яка була розроблена в 1990 році. Має наступні переваги:

— чистий синтаксис;
— відкритий код;
— зручна для розв’язання математичних проблем;
— має велику кількість модулів.

Недоліки:

— низька швидкодія;
— відсутність статичної типізації;

Наразі існує два діалекти (версії) Python: 3.7 та 3.6. Python 3 обернено не сумісний з попередньою серією 3. Код Python 2 швидше за все буде видавати помилки при виконанні в Python 3.

Нещодавно розробники оголосили про офіційне припинення розвитку гілки Python 3. Остання випущена версія Python 3.7. Далі розробка буде вестися лише у гілці Python 3.

Якщо задача, яку передбачається вирішити припускає наявність великої кількості даних, то, щоб уникнути помилок краще застосовувати 64-бітну версію мови програмування, тому що ця версія може отримати доступ до більш ніж 2 Гб пам'яті.

3.3.4.2 PyCharm Community Edition

PyChram – інтегроване середовище розробки для мови програмування Python [45], розроблене чеською компанією JetBrains. Можливості:

— статичний аналіз коду;
— навігація серед проектів;
— вбудовані інструменти юніт-тестування;
— підтримка систем контролю версій.

Існує дві загальні версії PyCharm: Professional Edition та Community Edition. Professional Edition має декілька варіантів ліцензій, які відрізняються функціональністю, вартістю та умовами використання. Community Edition безкоштовна версія з відкритим кодом, яка має усічений набір можливостей.

3.3.4.3 scikit-learn

Scikit-learn – безкоштовна бібліотека машинного навчання з відкритим кодом для мови Python [42, 39]. Вона має різні алгоритми класифікації, регресії та кластеризації. На основі цієї бібліотеки розроблені бібліотеки sklearn-tfidf та sklearn-deltatfidf, які слугують для обчислення ваг слів у документах.

3.3.4.4 openpyxl

Openpyxl бібліотека з відкритим кодом для мови Python, яка дозволяє зчитувати та записувати дані в форматах xlsx/xlsm/xltx/xltm [43]. Професійна підтримка openpyxl здійснюється Clark Consaulting & Research та Adimian.

3.3.5 Застосування дельта TF-IDF

Оскільки модуль модуль Extract N-Gram Features form Text не передбачає міри дельта TF-IDF, то його треба замінити на готові документи для навчальної гілки та гілки перевірки класифікатора.

Так як початковий набір даних розділяється на набір для навчання та для оцінки моделі, то у тренувальну модель в Azure Machine Learning Studio з експерименту з загальною кількістю 11794 текстів було додано два модулі Convert to CSV для подальшої роботи з ними (рис. 3.9)

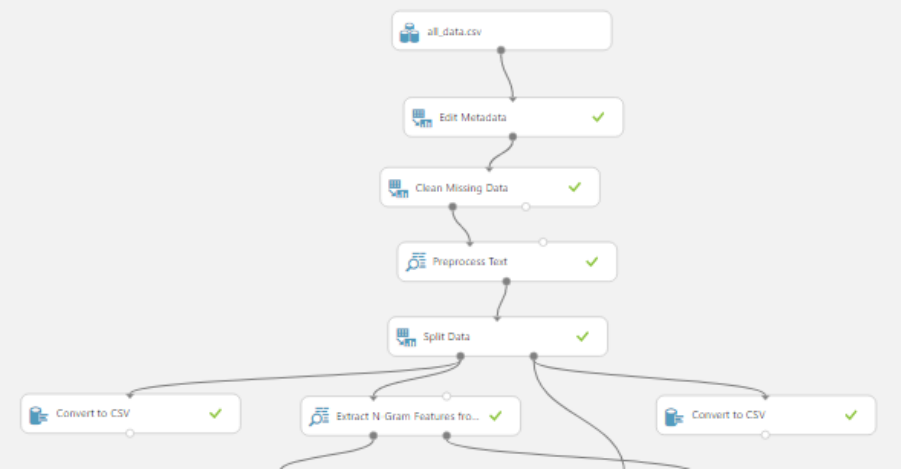

Рисунок 3.9 – Модернізована тренувальна модель

До моделі був доданий модуль Clean Missing Data, який видаляє відсутні значення.

Вихід 1 з модуля Split Data повертає дані для тренування моделі, а 2 – для її оцінки.

Весь скрипт розрахунку дельта TF-IDF показано в додатку Б.

Для роботи з бібліотекою openpyxl дані були конвертовані до формату xlsx. Для того, щоб завантажити файл для роботи були застосовані наступні команди:

from openpyxl import load_workbook

wb = load_workbook('C:/Users/Admin/Desktop/data_preprocessed (delta tfidf) train.xlsx')

sheet = wb.get_sheet_by_name("data")

Для перетворення тексту на вектори використовується бібліотека sklearn-deltatfidf, а саме модуль DeltaTfidfVectorizer. Цей модуль приймає багато параметрів, але для цього експерименту були використані наступні:

— ngram_range – нижня та верхня межа діапазону n-грам. У нашому випадку це 1 і 2, що означає що будуть виділені уніграми та біграми.

— max_features – максимальна кількість ознак (термів). У нашому випадку максимальна кількість дорівнює 1000.

— max_df – під час створення словника ігноруються терми, що мають частоту вище ніж задана. У нашому випадку це 0.8.

— min_df - під час створення словника ігноруються терми, що мають частоту нижче ніж задана. У нашому випадку це 5.

Усі параметри у модулі DeltaTfidfVectorizer дорівнюють параметрам з модулю Extract N Gram Features from Text у Azure Machine Learning Studio.

Щоб застосувати модуль використовується наступна команда (додаток В, рядок 23):

v = DeltaTFidfVectorizer(ngram_range=(1,2), max_features=1000, max_df=0.8, min_df=5)

Після цього застосовується метод fit_trnasform, що отримує документи та їх оцінки та повертає векторну модель (підрозділ 1.3.2) чи іншими словами матрицю з вагами термів. Оцінки мають мати наступний вигляд: 1 – позитивний документ; -1 – негативний; 0 – нейтральний. Оскільки тренувальний набір даних має оцінки вигляду positive, negative, neutral, їх треба перетворити. Зробити це можна за допомогою інструментів Microsoft Excel чи засобами мови Python як у нашому випадку (додаток В, рядки 31-54 та 71-75). Нижче наведено приклад команд.

```
d = list()
For I in range(2,5894): #5892 texts  from data, because row number starts from 1, first row is number of cols
        d.append(sheet.cell(row = i, column =5).value)
        if(sheet.cell(row = i, column = 4).value==”positive”:
                l.append(1)
        if(sheet.cell(row = i, column = 4).value==”negative”:
                l.append(-1)
        if(sheet.cell(row = i, column = 4).value==”neutral”:
                l.append(0)
a = v.fit_transform(d,l)
tf_idf_matr = a.todense()
zipped = tf_idf_matr.tolist()
```

Після цього отриману матрицю треба записати до файлу (додаток В, рядки 56 та 77-82). Це можна зробити наступним чином:

```
wb_new = openpyxl.Workbook()
```

```
sheet = wb_new.create_sheet("data")
c=1
for i in range(0,5892):
   for j in range(0,999):
      sheet.cell(row=c1, column=j+1).value = zipped[i][j]
   c=c+1
wb_new.save('C:/Users/Admin/Desktop/data_new.xlsx')
```

Цей блок коду може виконуватися довго, оскільки кількість значень, що записуються у файл – 5892000.

На етапі оцінки моделі в базовому експерименті використовується словник, отриманий з етапу навчання. У нашому випадку треба отримати позицію ознаки/терма в словнику, за цією позицією отримати значення IDF , отримати значення DF для позитивних та негативних документів – число документів у яких зустрічається ознака/терм. Словник отриманий за допомогою DeltaTfidfVectorizer зберігає значення у вигляді пари ключ-значення (рис. 3.10), де ключ – це ім’я ознаки/терма, а значення – позиція у словнику (списку значень idf) (рис. 3.11).

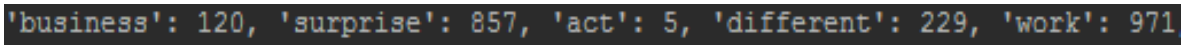

Рисунок 3.10 – Словник

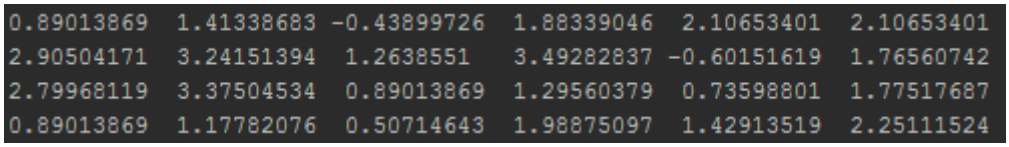

Рисунок 3.11 – Список значень IDF

Перевіримо значення отримані в підрозділі 3.3.2 для слова «comment». Щоб отримати позицію ознаки/терма у словнику треба виконати наступну команду:

```
term_n=v.vocabulary_.get('comment')
```

Для отримання значення IDF:

```
v.idf_[term_n]
```

Отримане значення 0,7757, що кардинально відрізняється від теоретичного отриманого в підрозділі 3.3.2.

Для обчислення IDF у бібліотеці sklearn-deltatfidf застосовується перша частина формули 3.11, але з деякими змінами:

$$IDF = \log\frac{|Pos|}{Pos_i} - \log\frac{|Neg|}{Neg_i} + 1 \quad (3.14)$$

де IDF – значення IDF у документі;

Оскільки для обчислення значення IDF не важливо яку основу буде мати логарифм (підрозділ 1.3.2), sklearn-deltatfidf використовує натуральний логарифм, Microsoft Azure Machine Learning Studio - десятковий. Також для значення IDF не потрібно знати частоту ознаки/терма у документі, тому у формулі воно відсутнє. До результату формули додають 1, щоб відкинути ознаки у яких IDF дорівнює 0. Також за замовчуванням параметр smoott_idf в модулі DeltaTfidfVectorizer дорівнює True. Smootth_idf робить припущення, що в корпусі документів є один документ, що включає до себе кожну ознаку/терм та додає 1 до кількості позитивних, негативних документів та до кількості документів у яких присутня ознака/терм.

Враховуючи написане вище підставимо нові значення у формулу 3.14.

$$IDF = \log\frac{2239}{37} - \log\frac{1666}{23} + 1 = 0{,}8202 \quad (3.15)$$

Нове значення наблизилось до фактичного.

Значення дельта TF-IDF для слова «comment» у 66 документі: 0,31354. Воно відрізняється від теоретичного, тому що в DeltaTfIdfVectorizer для нормалізації векторів застосовується L2-нормалізація (Евклідова норма) [32, 44], а в формулі 3.14 частота ознаки/терма обраховується за формулою 1.3 в якій кількість входжень ознаки в документ ділиться на загальну кількість слів документа, що є теж свого роду нормалізацією.

Далі треба розрахувати дельта TF-IDF для даних гілки перевірки класифікатора з урахуванням словника з навчальної гілки (додаток В, рядки 128-149). Приклад команд показано нижче.

```
v = DeltaTfidfVectorizer(vocabulary=v1.vocabulary_, max_df=0.8, min_df=1,
ngram_range=(1,2), max_features=1000) #v1.vocabulary_ vocabulary from train set
 p=v.fit_transform(d1,s)
tf_idf_matr = p[0].todense()
zipped = tf_idf_matr.tolist()
print(tf_idf_matr.shape)
 c=1
for i in range(0,5896):
    for j in range(0,999):
```

```
            sheet.cell(row=c, column=j+1).value = zipped1[i][j]
        wb_new.save('C:/Users/Admin/Desktop/score_dataset_deltatfidf.xlsx')
```

Для перевірки словника, який застосовується для оцінки моделі треба зробити наступне: щоб отримати необхідні значення DF треба внести деякі зміни до методу fit_transform бібліотеки DeltaTfidfVectorizer (додаток В, рядки 7-12), щоб він повертав значення X_pos та X_neg – кількість позитивних та негативних документів для окремої ознаки/терма та виконати запис у файл (додаток В, рядки 84-126). Блок-схема алгоритму наведена в додатку А, рис. А3. Приклад команд зазначено нижче.

```
def fit_transform(self, raw_documents, y):
   X, X_pos, X_neg = self._fit_transform(raw_documents, y)
    self._tfidf.fit(X_pos, X_neg, y)
    return self._tfidf.transform(X, copy=False), X_pos, X_neg
X_pos = document_frequency(a[1])+1
X_neg = document_frequency(a[2])+1
df=0
for i, f_name in enumerate(names):
   pos = v.vocabulary_[f_name]
   val = v.idf_[pos]
  for n in range(0, 5892):
              for j in range(0, len(zipped[n])):
                        if zipped[n][j] == val:
                                  df=df+1
   df_pos = X_pos[pos]
   df_neg = X_neg[pos]
   sheet_3.cell(row=i+2, column=1).value = i+1
   sheet_3.cell(row=i+2, column=2).value = f_name
  sheet_3.cell(row=i+2, column=3).value=df
   sheet_3.cell(row=i + 2, column=4).value = df_pos
   sheet_3.cell(row=i + 2, column=5).value = df_neg
   sheet_3.cell(row=i+2, column=6).value = val
  df=0
wb_new.save('C:/Users/Admin/Desktop/vocabulary_score(df pos neg).xlsx'
```

Далі була побудована тренувальна модель (рис. 3.12) в Azure Machine Learning Studio.

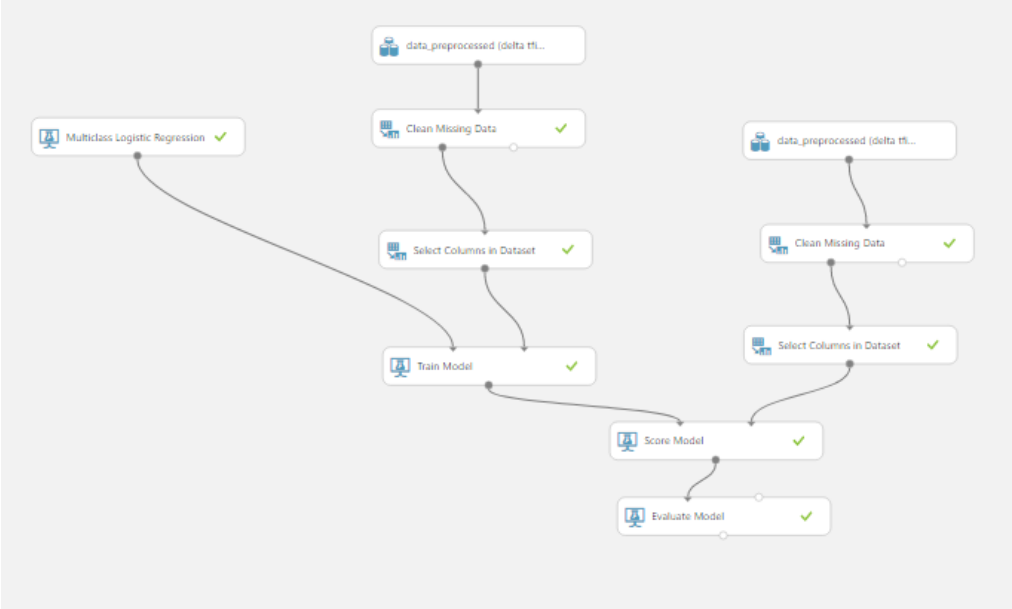

Рисунок 3.12 – Нова тренувальна модель

Були додані модулі Clean Missing Data після даних для тренування та після даних для оцінки моделі для видалення пустих значень.

Оцінкою моделі можу слугувати рисунок 3.13 на якому зображена матриця неточностей та таблиця 3.5.

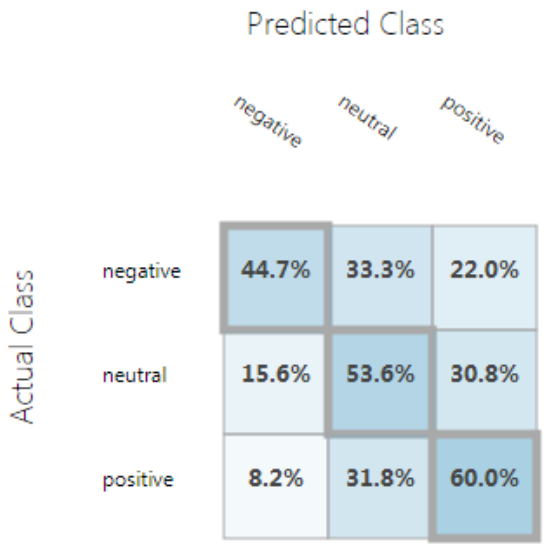

Рисунок 3.13 – Матриця неточностей для тренувальної моделі з дельта TF-IDF

Таблиця 3.5 – Метрики для оцінки моделі з експерименту з мірою дельта TF-IDF

Overall accuracy	Average accuracy	Micro-averaged precision	Macro-averaged precision	Micro-averaged recall	Macro-averaged recall
0,535787	0,690525	0.535758	0.544992	0.535787	0.527671

З матриці неточностей видно, що класифікатор правильно спрогнозував 44,7% негативних документів, 53,6% нейтральних та 60% позитивних.

Оскільки при використанні прогнозної моделі у нових даних немає оцінок, відповідно ми не знаємо кількість позитивних, негативних, нейтральних документів та кількість документів в яких з'являється та чи інша ознака, для розрахунку ваг для слів з нових даних використовується міра TF-IDF, оскільки недостатньо значень для обрахунку за формулою (3.14).

Застосування нової тренувальної моделі на нерозмічених даних дало результат показаний на рисунку 3.14.

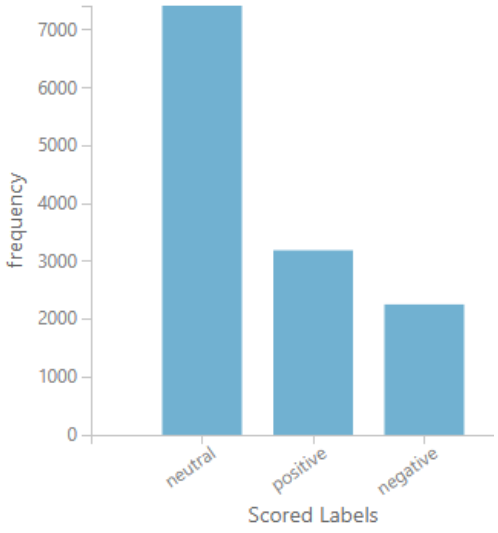

Рисунок 3.14 - Результат прогнозної моделі при використанні дельта TF-IDF

7423 (58%) елементів визначило як нейтральні, 3192 (25%) – позитивні, 1355 (17%) – негативні. Використання міри дельта TF-IDF дозволило зменшити кількість елементів, що відносяться до найбільшого класу на 23% порівняно з експериментом у якому використовувалася більша кількість даних та збільшити кількість елементів, що відносяться до позитивного та негативного класів (на 13% та 10.7% відповідно).

3.4 Висновки до третього розділу

В розділі «Методи оцінки моделей прогнозування та методи боротьби з незбалансованими даними» були:

— Розглянуті основні методи оцінки моделей прогнозування.

— Розглянуті методи боротьби з незбалансованим даними.

— Проведені експерименти для виявлення найкращого метода боротьби з незбалансованим даними.

Також була реалізована міра дельта TF-IDF за допомогою мови програмування Python та допоміжних бібліотек, оскільки модуль, який використовується для векторизації в Microsoft Azure Machine Learning Studio не передбачає її реалізації.

З проведених експериментів можна зробити висновок, що найефективнішим підходом методом боротьби з незбалансованими даними є застосування міри дельта TF-IDF замість TF-IDF.

4 ОХОРОНА ПРАЦІ ТА БЕЗПЕКА В НАДЗВИЧАЙНИХ СИТУАЦІЯХ. ЕКОЛОГІЯ

В даному розділі проведено аналіз потенційних небезпечних та шкідливих виробничих факторів, причин пожеж. Розглянуті заходи, які дозволяють забезпечити гігієну праці і виробничу санітарію. На підставі аналізу розроблені заходи з техніки безпеки та рекомендації з пожежної профілактики.

Завданням даної магістерської роботи було дослідити і виявити проблеми при роботі з даними з соціальних мереж та підвищити точність емоційної класифікації., і як результат було виявлено проблему незбалансованості даних та запропоновано способи її вирішення. В подальшому розроблятиметься реальна система, яка значно полегшить процес емоційної класифікації. Так як в процесі проектування використовувалося різне програмне забезпечення, то аналіз потенційно небезпечних і шкідливих виробничих чинників виконується для персонального комп'ютера.

4.1 Загальні питання з охорони праці

Умови праці на робочому місці, безпека технологічних процесів, машин, механізмів, устаткування та інших засобів виробництва, стан засобів колективного та індивідуального захисту, що використовуються працівником, а також санітарно-побутові умови повинні відповідати вимогам нормативних актів про охорону праці. В законі України «Про охорону праці» визначається, що охорона праці - це система правових, соціально-економічних, організаційно-технічних, санітарно-гігієнічних і лікувально-профілактичних заходів та засобів, спрямованих на збереження життя, здоров'я і працездатності людини у процесі трудової діяльності.

При роботі з обчислювальною технікою змінюються фізичні і хімічні фактори навколишнього середовища: виникає статична електрика, електромагнітне випромінювання, змінюється температура і вологість, рівень вміст кисню і озону в повітрі. Повітря забруднюється шкідливими хімічними речовинами антропогенного походження за рахунок деструкції полімерних матеріалів, які використовуються для обробки приміщень та обладнання. Неправильна організація робочого місця сприяє загальному і локальній напрузі м'язів шиї, тулуба, верхніх кінцівок, викривлення хребта і розвитку остеохондрозу. На всіх підприємствах, в установах, організаціях повинні створюватися безпечні і нешкідливі умови праці. Забезпечення цих умов покладається на власника або

уповноважений ним орган (далі роботодавець). Умови праці на робочому місці, безпека технологічних процесів, машин, механізмів, устаткування та інших засобів виробництва, стан засобів колективного та індивідуального захисту, що використовуються працівником, а також санітарно-побутові умови повинні відповідати вимогам нормативних актів про охорону праці. Роботодавець повинен впроваджувати сучасні засоби техніки безпеки, які запобігають виробничому травматизмові, і забезпечувати санітарно-гігієнічні умови, що запобігають виникненню професійних захворювань працівників. Він не має права вимагати від працівника виконання роботи, поєднаної з явною небезпекою для життя, а також в умовах, що не відповідають законодавству про охорону праці. Працівник має право відмовитися від дорученої роботи, якщо створилася виробнича ситуація, небезпечна для його життя чи здоров'я або людей, які його оточують, і навколишнього середовища.

4.1.1 Правові та організаційні основи охорони праці

Основним організаційним напрямом у здійсненні управління в сфері охорони праці є усвідомлення пріоритету безпеки праці і підвищення соціальної відповідальності держави, і особистої відповідальності працівників.

Державна політика в галузі охорони праці визначається відповідно до Конституції України Верховною Радою України і спрямована на створення належних, безпечних і здорових умов праці, запобігання нещасним випадкам та професійним захворюванням. Відповідно до статті 3 Закону України «Про охорону праці» (далі – Закону) законодавство про охорону праці складається з Закону, Кодексу законів про працю України, Закону України "Про загальнообов'язкове державне соціальне страхування від нещасного випадку на виробництві та професійного захворювання, які спричинили втрату працездатності" та прийнятих відповідно до них нормативно-правових актів, норм міжнародного договору (ратифіковані Конвенції і Рекомендації МОТ, директиви Європейської Ради).

На законодавчому рівні визначено такі пріоритетні напрямки з безпеки праці:

– кожен працівник несе безпосередню відповідальність за порушення зазначених Законом, нормами і правилами вимог;

– напрямки реалізації конституційного права громадян на їх життя і здоров’я в процесі трудової діяльності:

– пріоритет життя і здоров’я працівників по відношенню до результатів виробничої діяльності підприємства;

– повна відповідальність роботодавця за створення належних – безпечних і

здорових умов праці;

– соціальний захист працівників, повне відшкодування збитків особам, які потерпіли від нещасних випадків на виробництві та професійних захворювань;

– комплексне розв’язання завдань охорони праці;

– підвищення рівня промислової безпеки шляхом забезпечення суцільного технічного контролю за станом виробництв, технологій та продукції, а також сприяння підприємствам у створенні безпечних та нешкідливих умов праці;

– соціальний захист працівників, повне відшкодування збитків особам, які потерпіли від нещасних випадків на виробництві та професійних захворювань;

– використання економічних методів управління охороною праці, участь держави у фінансуванні заходів щодо охорони праці;

– використання світового досвіду організації роботи щодо поліпшення умов і підвищення безпеки праці на основі міжнародної співпраці.

Користувачі персональних комп’ютерів, для яких ця робота є головною, підлягають медичним оглядам: попереднім — під час влаштування на роботу і періодичним — протягом професійної діяльності раз на два роки.

4.1.2 Організаційно-технічні заходи з безпеки праці

В організації/підприємстві проводиться навчання і перевірка знань з питань охорони праці відповідно до вимог Типового положення про порядок проведення навчання і перевірки знань з питань охорони праці, затвердженого наказом Держнаглядохоронпраці України від 26.01.2005 N 15, зареєстрованого в Міністерстві юстиції України 15.02.2005 за N 231/10511 [46].

Також впроваджені організаційні заходи з пожежної безпеки - навчання і перевірку знань відповідно до вимог Типового положення про інструктажі, спеціальне навчання та перевірку знань з питань пожежної безпеки на підприємствах, в установах та організаціях України, затвердженого наказом Міністерства України з питань надзвичайних ситуацій та у справах захисту населення від наслідків Чорнобильської катастрофи від 29.09.2003 N 368, зареєстрованого в Міністерстві юстиції України 11.12.2003 за N 1148/8469 [47].

Обов’язковими вимогами враховане наступне:

– не слід допускати до роботи осіб, що в установленому порядку не пройшли навчання, інструктаж та перевірку знань з охорони праці, пожежної безпеки та цих Правил.

– на підприємстві/організації, де експлуатуються ЕОМ з відео дисплейними терміналами (ВДТ) і периферійними пристроями (ПП), розробляється інструкція з охорони праці відповідно до Положення про розробку інструкцій з охорони праці, затвердженого наказом Держнаглядохоронпраці від 29.01.98 N 9, зареєстрованого в Міністерстві юстиції України 07.04.98 за N 226/2666 [53].

– ознайомлення з правилами безпеки праці, одержання відповідних інструктажів засвідчується у журналі інструктажів.

– перед допуском до самостійної роботи кожен працівник має право на навчання з питань охорони праці і роботодавець зобов'язаний, і проводить таке навчання у вигляді двох інструктажів з питань охорони праці:

1) вступного, який проводять працівники служби охорони праці об'єкта господарювання з усіма працівниками, яких приймають на роботу незалежно від їхньої освіти та стажу роботи за програмою, в якій подають загальні питання охорони праці із врахуванням її особливостей на об'єкті господарювання;

2) первинного, який проводять керівники структурних підрозділів на місці праці з кожним працівником до початку їхньої роботи на цьому робочому місці.

Проходження працівником цих інструктажів з питань охорони праці підтверджується записами у відповідних журналах обліку інструктажів і скріплюється підписами осіб, які проводили інструктажі та осіб, які отримали інструктажі.

3) Повторний (не рідше одного разу в 6 місяців);

А) Позаплановий (при зміні правил охорони праці);

5) Поточний (проводять з працівниками перед виконанням робіт, на яких оформляється наряд-допуск)

– обов'язкові організаційні заходи перед початком, під час і після завершення роботи повинні включати перевірку (візуально) наявності і справності електрообладнання та його заземлення. Після закінчення роботи - вимагається прибирання робочого місця, відключення всіх електроприладів від електромережі.

Не допускається:

– виконувати обслуговування, ремонт та налагодження ЕОМ з ВДТ і ПП безпосередньо на робочому місці оператора;

– зберігати біля ЕОМ з ВДТ і ПП папір, інші носії інформації, запасні блоки, деталі тощо, якщо вони не використовуються для поточної роботи;

– відключати захисні пристрої, самочинно проводити зміни у конструкції та складі ЕОМ з ВДТ і ПП або їх технічне налагодження;

– працювати з ВДТ, у яких під час роботи з'являються нехарактерні сигнали, нестабільне зображення на екрані тощо;

4.2 Аналіз стану умов праці

Робота над створенням системи емоційної класифікації проходитиме в приміщенні багатоквартирного будинку на третьому поверсі. Для даної роботи достатньо однієї людини, для якої надано робоче місце зі стаціонарним комп’ютером.

4.2.1 Вимоги до приміщень

Геометричні розміри приміщення зазначені в табл. 4.1.

Таблиця 4.1 – Розміри приміщення

Найменування	**Значення**
Довжина, м	6
Ширина, м	2,5
Висота, м	2,4
Площа, $м^2$	15
Об’єм, $м^3$	36

Згідно з [48] розмір площі для одного робочого місця оператора персонального комп’ютера має бути не менше 6 кв. м, а об’єм — не менше 20 куб. м. Отже, дане приміщення цілком відповідає зазначеним нормам.

Для зручності роботи з в кімнаті є диван і журнальний стіл, обставлені живими квітами. Також робочий процес пов’язаний з багатьма документами, теками, журналами для чого приміщення облаштоване принтером і шафою для зручності. Задля дотримання визначеного рівня мікроклімату в будівлі встановлено систему опалення та кондиціонування.

Для забезпечення потрібного рівного освітленості кімната має вікно та систему загального рівномірного освітлення, що встановлена на стелі.

4.2.2 Вимоги до організації місця праці

При порівнянні відповідності характеристик робочого місця нормативним основні вимоги до організації робочого місця за [49] і відповідними фактичними значеннями для робочого місця, констатується повна відповідність.

Таблиця 4.2 - Характеристики робочого місця

Найменування параметра	Фактичне значення	Нормативне значення
Висота робочої поверхні, мм	750	680 ÷800
Висота простору для ніг, мм	730	не менше 600
Ширина простору для ніг, мм	660	не менше 500
Глибина простору для ніг, мм	700	не менше 650
Висота поверхні сидіння, мм	470	400 ÷500
Ширина сидіння, мм	400	не менше 400
Глибина сидіння, мм	400	не менше 400
Висота поверхні спинки, мм	600	не менше 300
Ширина опорної поверхні спинки, мм	500	не менше 380
Радіус кривини спинки в горизонтальній площині, мм	400	400
Відстань від очей до екрану дисплея, мм	800	700 ÷800

Робочий стіл на досліджуваному місці також містить достатньо простору для ніг. Крісло, що використовується в якості робочого сидіння, є підйомно-поворотним, має підлокітники і можливість регулювання за висотою і кутом нахилу спинки, також воно м'яке і виконане з екологічної шкіри, що дає можливість працювати у комфорті. Екран монітору знаходиться на відстані 0.8 м, клавіатура має можливість регулювання кута нахилу 5-15°. Отже, за всіма параметрами робоче місце відповідає нормативним вимогам.

Приміщення знаходиться на третьому поверсі п’ятиповерхової будівлі і має об’єм 36 м^3, площу – 15 м^2. У цьому кабінеті обладнано місце праці, яке укомплектовано ПК.

Температура в приміщенні протягом року коливається у межах 18–24°С, відносна вологість — близько 50%. Швидкість руху повітря не перевищує 0,2 м/с. Шум знаходиться на рівні 50 дБА. Система вентилювання приміщення — природна неорганізована, а опалення — індивідуальне.

Розміщення вікон забезпечує природне освітлення з коефіцієнтом природного освітлення не менше 1,5%, а загальне штучне освітлення, яке здійснюється за допомогою дев’яти люмінесцентних ламп, забезпечує рівень освітленості не менше 200 Лк.

У приміщенні є електрична мережа з напругою 220 В, яка створює небезпеку ураження електричним струмом. ПК та периферійні пристрої можуть бути джерелами

електромагнітних випромінювань, аерозолів та шкідливих речовин (часток тонеру, оксидів нітрогену та озону).

За ступенем пожежної безпеки приміщення належить до категорії В.

4.2.3 Навантаження та напруженість процесу праці

Як приклад наведено опис процесу праці ”*оформлення роботи”* під час виконання магістерської роботи: за фізичним навантаженням робота відноситься до категорії легкі роботи (Іа), її виконують сидячи з періодичним ходінням. Щодо характеру організування виконання дипломної роботи, то він підпадає під нав’язаний режим, оскільки певні розділи роботи необхідно виконати у встановлені конкретні терміни. За ступенем нервово-психічної напруги виконання роботи можна віднести до ІІ – ІІІ ступеня і кваліфікувати як помірно напружений – напружений за умови успішного виконання поставлених завдань.

Під час виконання робіт використовують ПК та периферійні пристрої (лазерні та струменеві), що призводить до навантаження на окремі системи організму. Такі перекоси у напруженні різних систем організму, що трапляються під час роботи з ПК, зокрема, значна напруженість зорового аналізатора і довготривале малорухоме положення перед екраном, не тільки не зменшують загального напруження, а навпаки, призводять до його посилення і появи стресових реакцій.

Тобто наявні психофізіологічні небезпечні та шкідливі фактори:

а) фізичного перевантаження:

- статичного;
- динамічного;

б) нервово-психічного перевантаження:

- розумового перенапруження;
- монотонності праці;
- перенапруження аналізаторів;
- емоційних перевантажень.

Рекомендовано застосування екранних фільтрів, локальних світлофільтрів (засобів індивідуального захисту очей) та інших засобів захисту, а також інші профілактичні заходи на ведені в [49]

Роботу за дипломним проектом визнано, таку, що займає 50% часу робочого дня та за восьмигодинної робочої зміни рекомендовано встановити додаткові регламентовані перерви:

- для розробників програм тривалістю 15 хв через кожну годину роботи;
- для операторів персональних комп’ютерів тривалістю 15 хв через дві години роботи;
- для операторів комп’ютерного набору тривалістю 10 хв через кожну годину роботи.

4.3 Виробнича санітарія

На підставі аналізу небезпечних та шкідливих факторів при виробництві (експлуатації), пожежної безпеки можуть бути надалі вирішені питання необхідності забезпечення працюючих достатньою кількістю освітлення, вентиляції повітря, організації заземлення, тощо.

4.3.1 Аналіз небезпечних та шкідливих факторів при виробництві (експлуатації) виробу

Роботу, пов'язану з ЕОМ, виконують із забезпеченням виконання [54] «Правил охорони праці під час експлуатації електронно-обчислювальних машин», які встановлюють вимоги безпеки до обладнання робочих місць, до роботи із застосуванням ЕОМ..

Основними робочими характеристиками персонального комп'ютера є наступні:

– робоча напруга U = +220В ± 5%;
– робочий струм I = 2А;
– споживана потужність P = 350 Вт.

Робоче місце має відповідати вимогам Державних санітарних правил і норм роботи з візуальними дисплейними терміналами електронно-обчислювальних машин, затверджених постановою Головного державного санітарного лікаря України від 10.12.98 N 7 [49].

За умов роботи з ПК виникають наступні небезпечні та шкідливі чинники: несприятливі мікрокліматичні умови, освітлення, електромагнітні випромінювання, забруднення повітря шкідливими речовинами (джерелом, яких можуть бути: принтер, сканер та інші джерела виділення багатьох хімічних речовин - напр., озону, оксидів азоту та аерозолів високодисперсних частинок тонера), шум, вібрація, електричний струм, електростатичне поле, напруженість трудового процесу та інше.

Аналіз небезпечних та шкідливих виробничих факторів виконується у табличній формі (табл. Г.1 додатку Г).

4.3.2 Пожежна безпека

Небезпека розвитку пожежі обумовлюється застосуванням розгалужених систем електроживлення ЕОМ, вентиляції і кондиціювання,. Небезпека загоряння пов'язана з особливістю комп'ютерів - із значною кількістю щільно розташованих на монтажній платі і блоках електронних вузлів і схем, електричних і комутаційних кабелів, резисторів, конденсаторів, напівпровідникових діодів і транзисторів. Надійна робота окремих елементів і мікросхем в цілому забезпечується тільки в певних інтервалах температури, вологості і при заданих електричних параметрах. При відхиленні реальних умов експлуатації від розрахункових можуть виникнути пожежонебезпечні ситуації.

Висока щільність елементів в електронних схемах призводить до значного підвищення температури окремих вузлів (80…100 ℃). При проходженні електричного струму по провідниках і деталей виділяється тепло, що в умовах їх високої щільності може привести до перегріву, і може служити причиною запалювання ізоляційних матеріалів. Також ймовірна небезпека внаслідок перевантаження напруги, розрядки зарядів статичної електрики, пошкодження обладнання та електропроводки. Електростатичний розряд виникає під час тертя двох ізольованих матеріалів. Розряд статичної електрики може виникнути під час роботи вентилятора або комп'ютер. Кабельні лінії є найбільш пожежонебезпечними місцем. Наявність пального ізоляційного матеріалу, ймовірних джерел запалювання у вигляді електричних іскор і дуг, розгалуженість і недоступність роблять кабельні лінії місцем найбільш ймовірного виникнення і розвитку пожежі. Для зниження займистості і здатності поширювати полум'я кабелі покривають вогнезахисними покриттями.

Для гасіння пожеж пропонується використовувати порошкові або вуглекислотні вогнегасники, так як вони є універсальними. Заземлені конструкції, що знаходяться в приміщеннях, де розміщені робочі місця (батареї опалення, водопровідні труби, кабелі із заземленим відкритим екраном), надійно захищені діелектричними щитками та/або сітками з метою недопущення потрапляння працівника під напругу. Проходи до засобів пожежогасіння вільні, не захаращуються та у разі потреби забезпечувати евакуацію всіх людей, які перебувають у приміщенні через один евакуаційний вихід з дверима на шляху евакуації, що відчинятися в напрямку виходу з будівлі від робочого місця.

Запобігти утворенню горючого середовища (замінити горючі речовини і матеріали

на негорючі і важкогорючі) не надається технічно можливим. Тому проектом передбачаються способи і засоби запобігання утворення (або внесення) в горюче середовище джерел запалювання, таких як:

1) застосування електроустаткування, відповідної пожежонебезпечної і вибухонебезпечної зонам відповідно до ПУЕ;

2) застосування в конструкції швидкодійних засобів захисного відключення можливих джерел запалення;

3) виключення можливості появи іскрового розряду в горючому середовищі з енергією, рівної і вище мінімальної енергії запалення.

Згідно [50] таке приміщення, площею 15 м^2, відноситься до категорії "В" (пожежонебезпечної).

Виникнення пожежі можливе, якщо на об'єкті є горючі речовини, окислювач і джерела запалювання. Вірогідність пожежної небезпеки приймається значною, якщо ймовірна взаємодія цих трьох чинників.

Горючими матеріалами в приміщенні, де розташовані ЕОМ, є:

1) поліамід – матеріал корпусу мікросхем, горюча речовина, температура самозаймання 420º С,

2) полівінілхлорид – ізоляційний матеріал, горюча речовина, температура запалювання 335º С, температура самозаймання 530º С,

3) склотекстоліт ДЦ – матеріал друкарських плат, важкогорючий матеріал, показник горючості 1.74, не схильний до температурного самозаймання,

А) пластикат кабельний №.489 – матеріал ізоляції кабелів, горючий матеріал, показник горючості більше 2.1,

5) деревина – будівельний і обробний матеріал, з якого виготовлені меблі, горючий матеріал, показник горючості більше 2.1, температура запалювання 255º С, температура самозаймання 399º С.

Простори усередині приміщень в межах, яких можуть утворюватися або знаходиться пожежонебезпечні речовини і матеріали відповідно до [50] відносяться до пожежонебезпечної зони класу П-IIа. Це обумовлено тим, що в приміщенні знаходяться тверді горючі та важкозаймисті речовини та матеріали. Приміщенню, у якому розташоване робоче місце, присвоюється II ступень вогнестійкості.

Потенційними джерелами запалювання можуть бути:

1) іскри і дуги короткого замикання;

2) електрична іскра при замиканні і розмиканні ланцюгів;

3) перегріви від тривалого перевантаження,

4) відкритий вогонь і продукти горіння,

5) наявність речовин, нагрітих вище за температуру самозаймання,

6) розрядна статична електрика.

Причинами можливого загоряння і пожежі можуть бути:

1) несправність електроустановки;

2) конструктивні недоліки устаткування;

3) коротке замикання в електричних мережах;

4) запалювання горючих матеріалів, що знаходяться в безпосередній близькості від електроустановки.

Продуктами згорання, що виділяються на пожежі, є: окис вуглецю; сірчистий газ; окис азоту; синильна кислота; акромін; фосген; хлор і ін. При горінні пластмас, окрім звичних продуктів згорання, виділяються різні продукти термічного розкладання: хлорангідридні кислоти, формальдегіди, хлористий водень, фосген, синильна кислота, аміак, фенол, ацетон, стирол.

4.3.3 Електробезпека

На робочому місці виконуються наступні вимоги електробезпеки: ПК, периферійні пристрої та устаткування для обслуговування, електропроводи і кабелі за виконанням та ступенем захисту відповідають класу зони за ПУЕ (правила улаштування електроустановок), мають апаратуру захисту від струму короткого замикання та інших аварійних режимів. Лінія електромережі для живлення ПК, периферійних пристроїв і устаткування для обслуговування, виконана як окрема групова три- провідна мережа, шляхом прокладання фазового, нульового робочого та нульового захисного провідників. Нульовий захисний провідник використовується для заземлення (занулення) електроприймачів. Штепсельні з'єднання та електророзетки крім контактів фазового та нульового робочого провідників мають спеціальні контакти для підключення нульового захисного провідника.

4.4 Гігієнічні вимоги до параметрів виробничого середовища

4.4.1 Мікроклімат

Мікроклімат робочих приміщень – це клімат внутрішнього середовища цих приміщень, що визначається діючої на організм людини з'єднанням температури, вологості, швидкості переміщення повітря. В даному приміщенні проводяться роботи, що

виконуються сидячи і не потребують динамічного фізичного напруження, то для нього відповідає категорія робіт Іа. Отже оптимальні значення для температури, відносної вологості й рухливості повітря для зазначеного робочого місця відповідають [51] і наведені в табл. 4.3:

Таблиця 4.3 – Норми мікроклімату робочої зони об’єкту

Період року	**Категорія робіт**	**Температура C^0**	**Відносна вологість %**	**Швидкість руху повітря, м/с**
Холодна	легка-1 а	22 - 24	40 – 60	0,1
Тепла	легка-1 а	23 - 25	40 – 60	0,1

Дане приміщення обладнане системами опалення, кондиціонування повітря або припливно-витяжною вентиляцією. У приміщенні на робочому місці забезпечуються оптимальні значення параметрів мікроклімату: температури, відносної вологості й рухливості повітря у відповідності до [51]. Для забезпечення оптимальних параметрів мікроклімату в приміщенні проводяться перерви в роботі, з метою його провітрювання. Контроль параметрів мікроклімату в холодний і теплий період року здійснюється не менше 3-х разів на зміну (на початку, середині, в кінці).

4.4.2 Освітлення

Світло є природною умовою існування людини. Воно впливає на стан вищих психічних функцій і фізіологічні процеси в організмі. Хороше освітлення діє тонізуюче, створює гарний настрій, покращує протікання основних процесів вищої нервової діяльності.

Збільшення освітленості сприяє поліпшенню працездатності навіть в тих випадках, коли процес праці практично не залежить від зорового сприйняття. При поганому освітленні людина швидко втомлюється, працює менш продуктивно, виникає потенційна небезпека помилкових дій і нещасних випадків.

Освітленість приміщення має велике значення при роботі на ПЕОМ. Вона багато в чому визначається колірною і мережевою обстановкою. Для забарвлення стіни рекомендується віддавати перевагу світлим фарбам.

Основний потік природного світла при цій повинен бути зліва. Не допускається спрямування основного світлового потоку природного світла праворуч, ззаду і спереду

працівника на ПЕОМ.

Робота на ПЕОМ може здійснюватися за таких видах освітлення:

- загальному штучному освітленні, коли монітори розташовуються по периметру приміщення або при центральному розташуванні робочих місць у два ряди по довжині кімнати з екранами, звернені в протилежні сторони;

- суміщене освітлення (природне + штучне) тільки при одному і трьох рядном розташуванні робочих місць, коли екран і поверхню робочого столу знаходяться перпендикулярно світла несучій стіні. При цьому штучне освітлення буде виконане стельовими або підвісними люмінесцентними світильниками, рівномірно розміщеними по стелі рядами паралельно світловим прорізам так, щоб екран відео монітора знаходився в зоні захисного кута світильника, і його проекції не доводилися на екран. Працюючі на ПЕОМ не повинні бачити відображення світильників на екрані. Застосовувати місцеве освітлення при роботі на ПЕОМ не рекомендується.

Природне освітлення застосовується, коли робочі місця з ПЕОМ розташовуються в один ряд по довжині приміщення на відстані 0,8 - 1,0 м від стіни з віконними прорізами, і екрани знаходяться перпендикулярно цієї стіни. Основний потік природного світла повинен бути зліва. Не допускається спрямування основного світлового потоку природного світла праворуч, ззаду і спереду працює на ПЕОМ. Оптимальна відстань очей до екрана відео монітора повинна становити 60-70 см, допустиме не менше 50 см. Розглядати інформацію ближче 50 см не рекомендується.

У проекті, що розробляється, передбачається використовувати суміщене освітлення. У світлий час доби використовуватиметься природне освітлення приміщення через віконні отвори, в решту часу використовуватиметься штучне освітлення. Штучне освітлення створюється газорозрядними лампами.

Штучне освітлення в робочому приміщенні передбачається здійснювати з використанням люмінесцентних джерел світла в світильниках загального освітлення. Люмінесцентні лампи мають потужність 15 Вт, тривалий термін служби (до 10000 годин). При експлуатації ЕОМ виконується зорова робота IV в розряду точності (середня точність). При цьому нормована освітленість на робочому місці (Ен) рівна 200 лк. Джерелом природного освітлення є сонячне світло.

У приміщенні, де розташовані ЕОМ передбачається природне бічне освітлення, рівень якого відповідає [55]. Джерелом природного освітлення є сонячне світло. Регулярно повинен проводитися контроль освітленості, який підтверджує, що рівень освітленості задовольняє ДБН і для даного приміщення в світлий час доби достатньо природного освітлення.

Розрахунок освітлення.

Для даного приміщення світловий коефіцієнт приймається в межах 1/8 - 1/10:

$$\sqrt{a^2+b^2}\cdot S_b = (1/8 \div 1/10)\cdot S_n \quad (4.1)$$

де S_b – площа віконних прорізів, м^2;

S_n – площа підлоги, м^2.

$S_n = a \cdot b = 6 \times 2{,}5 = 15$ м^2

$S_{\text{вік}} = \frac{1}{8} \cdot 15 = 1{,}875$ м^2

Приймаємо 1 вікно площею $S = 1{,}8$ м^2.

Світильники загального освітлення розташовуються над робочими поверхнями в рівномірно-прямокутному порядку. Для організації освітлення в темний час доби передбачається обладнати приміщення, довжина якого складає 6 м, ширина 2,5 м, світильниками, люмінесцентними лампами типу (три по 25 Вт) з світловим потоком 1600 лм кожна.

Розрахунок штучного освітлення виробляється по коефіцієнтах використання світлового потоку, яким визначається потік, необхідний для створення заданої освітленості при загальному рівномірному освітленні. Розрахунок кількості світильників *n* виробляється по формулі (4.2):

$$n = \frac{E \cdot S \cdot Z \cdot K}{F \cdot U \cdot M} \quad (4.2)$$

де *E* – нормована освітленість робочої поверхні, визначається нормами – 300 лк;

S – освітлювана площа, м^2; S = 15 м^2;

Z – поправочний коефіцієнт світильника (Z = 1,15 для ламп розжарювання та ДРЛ; Z = 1,1 для люмінесцентних ламп) приймаємо рівним 1,1;

K – коефіцієнт запасу, що враховує зниження освітленості в процесі експлуатації – 1,5;

U – коефіцієнт використання, залежний від типу світильника, показника індексу приміщення і т.п. – 0,575

M – число ламп в світильнику – 3;

F – світловий потік лампи – 1600 лм.

Підставивши числові значення у формулу (4.2), отримуємо:

$$n = \frac{300 \cdot 15 \cdot 1{,}1 \cdot 1{,}5}{1600 \cdot 0{,}575 \cdot 3} \approx 3$$

Приймаємо освітлювальну установку, яка складається з 3-х світильників, які складаються з трьох люмінесцентних ламп загальною потужністю 75 Вт, напругою – 220 В.

4.5 Шум та вібрація, електромагнітне випромінювання

Рівень шуму, що супроводжує роботу користувачів персональних комп'ютерів (зумовлений як роботою системних блоків, клавіатури, так і друкуванням на принтерах, а також зовнішніми чинниками), коливається у межах 50–65 дБА [52]. Шум такої інтенсивності на тлі високого ступеня напруженості праці негативно впливає на функціональний стан користувачів. Тому на практиці рекомендують знижувати фактичний рівень шуму у приміщеннях, де створюють комп'ютерні програми, виконують теоретичні та творчі роботи, проводять навчання до 40 дБА, а в приміщеннях, де виконують роботу, що потребує зосередженості, — до 55 дБА.

Шум часто є причиною зниження рівня працездатності, підвищення рівня загальної та професійної захворюваності, частоти виробничих травм. Шум є загальнобіологічним подразником, який негативно впливає на всі органи і системи організму. У разі тривалого систематичного впливу шуму може виникнути патологія з переважним ураженням слуху, центральної нервової і серцево- судинної систем.

У приміщенні з ЕОМ коректований рівень звукової потужності не перевищує 45 дБА. Оскільки рівень шуму не перевищує гранично допустимих величин, які встановлені санітарними нормами, заходи для зниження шуму не проводяться.

Вібрація на робочому місці в приміщенні, що розглядається, відповідає нормам [52]. Допустимий рівень вібрацій на робочому місці: для 1 ступеня шкідливості до 3 дБ; для 2-3 - 1-6 дБ; для 3 - більше 6 дБ.

Для захисту від електромагнітного випромінювання передбачаються наступні заходи:

1) застосування нових моніторів,

2) віддалення робочого місця не менше, ніж на 0,4-0,5 м, оскільки напруженість електричного поля зменшується при віддаленні від джерела поля,

3) встановлення раціональних режимів роботи персоналу (обмеження часу перебування),

4) раціональне розміщення в робочому приміщенні устаткування, що випромінює електромагнітну енергію.

4.6 Вентилювання

У приміщенні, де знаходяться ЕОМ, повітрообмін реалізується за допомогою природної організованої вентиляції. Цей метод забезпечує приплив потрібної кількості свіжого повітря, що визначається в СНіП (30 м^3 на годину на одного працюючого).

Також має здійснюватися провітрювання приміщення, в залежності від погодних умов, тривалість повинна бути не менше 10 хв. Найкращий обмін повітря здійснюється при наскрізному провітрюванні.

4.7 Заходи з організації виробничого середовища та попередження виникнення надзвичайних ситуацій

Відповідно до санітарно-гігієнічних нормативів та правил експлуатації обладнання наводимо приклади деяких заходів безпеки.

1) Заходи безпеки під час експлуатації персонального комп'ютера та периферійних пристроїв передбачають:

- правильне організування місця праці та дотримання оптимальних режимів праці та відпочинку під час роботи з ПК;
- експлуатацію сертифікованого обладнання;
- дотримання заходів електробезпеки;
- забезпечення оптимальних параметрів мікроклімату;
- забезпечення раціонального освітлення місця праці (освітленість робочого місця не перевищувала 2/3 нормальної освітленості приміщення);
- облаштовуючи приміщення для роботи з ПК, потрібно передбачити припливно-витяжну вентиляцію або кондиціювання повітря:

а) якщо об'єм приміщення 20 м^3, то потрібно подати не менш як 30 м^3/год повітря;

б) якщо об'єм приміщення у межах від 20 до 40 м^3, то потрібно подати не менш як 20 м^3/год повітря;

в) якщо об'єм приміщення становить понад 40 м^3, допускається природна вентиляція, у випадку, коли немає виділення шкідливих речовин.

- зниження рівня шуму та вібрації:

а) у джерелі виникнення, шляхом застосування раціональних конструкцій, нових матеріалів і технологічних процесів;

б) звукоізоляція устаткування за допомогою глушників, резонаторів, кожухів, захисних конструкцій, оздоблення стін, стелі, підлоги тощо;

в) використання засобів індивідуального захисту.

2) Заходи безпеки під час експлуатації інших електричних приладів передбачають дотримання таких правил:

- постійно стежити за справним станом електромережі, розподільних щитків, вимикачів, штепсельних розеток, лампових патронів, а також мережевих кабелів живлення, за допомогою яких електроприлади під'єднують до електромережі;

- постійно стежити за справністю ізоляції електромережі та мережевих кабелів, не допускаючи їхньої експлуатації з пошкодженою ізоляцією;

- не тягнути за мережевий кабель, щоб витягти вилку з розетки;

- не закривати меблями, різноманітним інвентарем вимикачі, штепсельні розетки;

- не підключати одночасно декілька потужних електропристроїв до однієї розетки, що може викликати надмірне нагрівання провідників, руйнування їхньої ізоляції, розплавлення і загоряння полімерних матеріалів;

- не залишати включені електроприлади без нагляду;

- не допускати потрапляння всередину електроприладів крізь вентиляційні отвори рідин або металевих предметів, а також не закривати їх та підтримувати в належній чистоті, щоб уникнути перегрівання та займання приладу;

- не ставити на електроприлади матеріали, які можуть під дією теплоти, що виділяється, спалахнути (канцелярські товари, сувенірну продукцію тощо).

Вимоги безпеки при надзвичайних ситуаціях:

1) При раптовому припиненні подачі електричної енергії вимкнути всі пристрої ПК в такій послідовності: периферійні пристрої, системний блок, стабілізатор (або блок безперервного живлення). Витягнути вилки з розеток. При наявності ознак горіння (дим, запах горілого) необхідно вимкнути всі пристрої ПК, знайти місце загоряння і виконати всі можливі заходи для його ліквідації, попередивши терміново про це керівництво. У випадку виникнення пожежі негайно попередити про це пожежну частину та керівництво, виконати усі можливі заходи по евакуації людей з приміщення і розпочати гасіння пожежі первинними засобами пожежогасіння.

2) При замиканні, перевантаженні електричного струму на електричному обладнанні, внаслідок ураження грозової блискавки та ймовірної небезпеки ураженням електричним струмом, приймають наступне:

- попередження замикання здійснюється правильним вибором, монтажем експлуатації мереж;

- застосування захисту схем у вигляді швидкодіючих реле, а також вимикачів, плавких запобіжників, автоматичних вимикачів.

а) У випадку дотику до корпусу та інших струмоведучих частин електроустановки, що опинилися під напругою використовують захисне заземлення - зниження до безпечних значень напруги дотику і кроку, обумовлених замиканням на корпус та ін. Це досягається шляхом, зменшення потенціалу заземленого обладнання (за рахунок підйому потенціалу підстави, на якому стоїть людина, до значення, близького до значення потенціалу заземленого обладнання) та відключення від загальної електромережі ураженого обладнання.

б) У випадку замикання фази на корпус, зниження ізоляції мережі нижче визначеної межі і, нарешті, в разі дотику людини безпосередньо до частини, що знаходиться під напругою. Основними елементами пристрою захисного відключення є прилад захисного відключення і автоматичний вимикач.

Прилад захисного відключення - сукупність окремих елементів, які приймають вхідну величину, реагує на її зміни і при заданому значенні дають сигнал на її відключення вимикача:

- датчику - вхідна ланка пристрою, що сприймають впливу ззовні і здійснюють перетворення цього впливу в відповідний сигнал;

- підсилювача, призначений для посилення сигналу датчика, якщо він виявляється недостатньо потужним;

- ланцюгів контролю, службовці періодичної перевірки справності захисного відключення;

- допоміжних елементів - сигнальні лампи і вимірювальні прилади, що характеризують стан електроустановки.

Автоматичний вимикач - апарат, призначений для включення і вимикання від ланцюгів під навантаженням і при коротких замиканнях. Він повинен включати ланцюг автоматично при надходженні сигналу від приладу захисного відключення.

Також застосовують різні **електричні захисні засоби від ураження струмом**:

а) Ізолюючі - ізолюють людини від струмоведучих або заземлених частин, а так-же від землі. Вони діляться на основні та додаткові.

б) Основні - володіють ізоляцією, здатної довго витримувати робоче напругу електроустановки і тому ними дозволяється стосуватися струмоведучих частин, знаходячи-трудящих під напругою. До них відносяться: в електроустановках до 1000 Вт - діелектричної рукавички, ізолюючі штанги, ізолюючі і електровимірювальні кліщі і т.д .; понад 1000Вт - ізолюючі штанги, і електровимірювальні кліщі, а також кошти для ремонтних робіт під напругою понад 1000Вт.

в) Запобіжні - володіють ізоляцією нездатною витримати робоча напруга електроустановки, і тому вони не можуть самостійно захищати людину від ураження струмом під цим напругою. Їх значення - посилити захисні дії основних і ізолюючих засобів, разом з якими вони повинні застосовуватися, при чому при використання основних захисних засобів достатньо застосування одного запобіжного захисного засобу. До запобіжних відносяться засоби в електроустановках до 1000Вт - діелектричні калоші килимки, а також ізолюючі підставки.

Розрахунок захисного заземлення (забезпечення електробезпеки будівлі).

Загальний опір захисного заземлення визначається за формулою:

$$R_{ззп} = \frac{R_з \cdot R_п}{R_п \cdot n \cdot \eta_з + R_з \cdot \eta_п}, \tag{4.3}$$

де Rз - опір заземлення, Ом;

Rш - опір опори, яке з'єднує заземлювачі, Ом;

п - кількість заземлювачів;

ηз - коефіцієнт екранування заземлювача; приймається в межах 0,2 ÷ 0,9; ηз = 0,7

ηш - коефіцієнт екранування сполучної стійки; приймається в межах 0,1 ÷ 0,7; ηш = 0,5;

Опір заземлення визначається за формулою:

$$R_з = \frac{\rho}{2\pi \cdot l} \cdot \left(\ln \frac{2 \cdot l}{d} + \frac{1}{2} \ln \frac{4 \cdot t + l}{4 \cdot t - l} \right), \tag{4.4}$$

де ρ - питомий опір ґрунту, залежить від типу грунту, Ом·м;

для піску - 400 ÷ 700 Ом·м; приймаємо ρ = 400 Ом·м;

l - довжина заземлювача, м; для труб - 2-3 м; l = 3 м;

d - діаметр заземлювача, м; для труб - 0,03-0,05 м; d = 0,05 м;

t - відстань від середини забитого в ґрунт заземлювача до рівня землі, м; t = 2 м.

$$R_з = \frac{400}{2 \cdot 3{,}14 \cdot 3}\left(\ln\frac{2 \cdot 3}{0{,}05} + \frac{1}{2}\ln\frac{4 \cdot 2 + 3}{4 \cdot 2 - 3}\right) = 110, \text{ Ом}$$

Опір смуги, що з'єднує заземлювачі, визначається за формулою:

$$R_ш = \frac{\rho}{2\pi \cdot L} \cdot \ln\frac{2 \cdot L^2}{b \cdot t^1}, \quad (4.5)$$

де L - довжина смуги, що з'єднує заземлювачі (м) і приблизно дорівнює периметру будівлі: $П_{буд.} = 42 \cdot 2 + 38 \cdot 2 = 160$ м; L = 160 м;

b - ширина смуги, м; b = 0,03 м;

t_1 - глибина заземлення від рівня землі, м; $t_1 = 0{,}5$ м.

$$R_n = \frac{400}{2 \cdot 3{,}14 \cdot 160} \cdot \ln\frac{2 \cdot 160^2}{0{,}03 \cdot 0{,}5} = 5{,}99, \text{ Ом}$$

Кількість заземлювачів захисного заземлення визначається за формулою:

$$n = \frac{2 \cdot R_з}{4 \cdot \eta_з}, \quad (4.6)$$

де 4 - допустимий загальний опір, Ом;

2 - коефіцієнт сезонності.

Визначаємо загальний опір захисного заземлення:

$$R_{ззп} = \frac{110 \cdot 5{,}99}{5{,}99 \cdot 79 \cdot 0{,}7 + 110 \cdot 0{,}5} = 1{,}7 \text{ Ом}$$

Висновок: дане захисне заземлення буде забезпечувати електробезпеку будівлі, так як виконується умова: $R_{ззп} < 4$ Ом.

3) При виникненню пожеж при роботі на ПЕОМ від таких можливими джерел запалювання як:

– іскри і дуги коротких замикань;

– перегрів провідників, резисторів та інших радіодеталей ПЕОМ, від тривалої перевантаження та наявність перехідного опору;

– іскри при розмиканні і розмиканні ланцюгів;

– розряди статичної електрики;

– необережному поводженню з вогнем, а також вибухи газо-повітряних і паро-повітряних сумішей.

В приміщенні не повинно накопичуватися сміття, непотрібний папір, мотлох та ін. речі, які не використовуються у виробничому процесі. У разі виникнення пожежі необхідно повідомити в найближчу пожежну частину, убезпечити інших працівників і по можливості прийняти кроки по запобіганню можливих наслідків та усуненню пожежі.

4.8 Охорона навколишнього природного середовища

4.8.1 Загальні дані з охорони навколишнього природного середовища

Діяльність за темою магістерської роботи, а саме: виявлення проблем при роботі з даними з соціальних мереж та підвищення точності емоційної класифікації, процес виконання якої впливає на навколишнє природнє середовище і регламентується нормами діючого законодавства: Законом України «Про охорону навколишнього природного середовища», Законом України «Про забезпечення санітарного та епідемічного благополуччя населення», Законом України «Про відходи», Законом України «Про охорону атмосферного повітря», Законом України «Про захист населення і територій від надзвичайних ситуацій техногенного та природного характеру», Водний кодекс України.

Основним екологічним аспектом в процесі діяльності за даними спеціальностями є процеси впливу на атмосферне повітря та процеси поводження з відходами, які утворюються, збираються, розміщуються, передаються на знешкодження, утилізацію, тощо в ІТ галузі.

В процесі діяльності виявлення проблем при роботі з даними з соціальних мереж та підвищення точності емоційної класифікації виникають процеси поводження з відходами ІТ галузі. Нижче надано перелік відходів, що утворюються в процесі роботи:

— Відпрацьовані люмінесцентні лампи - І клас небезпеки

— Батарейки та акумулятори (малі) -ІІІ клас небезпеки

— Акумулятор для джерел безперебійного живлення -ІІІ клас небезпеки

— Відходи друкуючих пристроїв - IV клас небезпеки

— Макулатура - IV клас небезпеки

— Матеріали пакувальні пластмасові забруднені (ємності з-під тонеру, фарби, інш.) - IV клас небезпеки

— Побутові відходи - IV клас небезпеки

4.8.2 Вимоги до збору, пакування та розміщення відходів ІТ галузі

Наводяться вимоги зберігання виявлених за своєю роботою відходів відповідно до вимог Державних санітарних правил і норм [56].

Відходи в міру їх накопичення збирають у тару, відповідну класу небезпеки, з дотриманням правил безпеки, після чого доставляють до місця тимчасового зберігання відходів відповідно до затвердженої схеми їх розміщення. Зазначені для зберігання відходів місця чи об'єкти повинні використовуватися лише для заявлених відходів.

Не допускається зберігання відходів у невстановлених схемою місцях, а також перевищення норм тимчасового зберігання відходів.

Способи тимчасового зберігання відходів визначаються видом, агрегатним станом і класом небезпеки відходів:

- Відходи I класу небезпеки зберігаються в герметичній тарі (сталеві бочки, контейнери). У міру наповнення тару з відходами закривають герметично сталевий кришкою;

- Відходи II класу небезпеки в залежності від агрегатного стану зберігаються в поліетиленових мішках, бочках, сховищах та інших видах тари, яка запобігає поширенню шкідливих речовин;

- Відходи III класу небезпеки зберігаються в тарі, яка забезпечує локалізацію зберігання, дозволяє виконувати вантажно-розвантажувальні і транспортні роботи і виключає поширення в ОС шкідливих речовин;

- Відходи IV класу небезпеки можуть зберігатися відкрито на промисловому майданчику у вигляді конусоподібної купи, звідки їх автонавантажувачем перевантажують у самоскид і доставляють на місце утилізації або захоронення;

Не допускається змішування відходів різних видів і класів небезпеки з будівельними і побутовими відходами, відходами дерев'яної, металевої, синтетичної тари, відходами текстильних матеріалів (старий спецодяг, ганчірки) і ін.

Особливий контроль наділяється збору і зберіганню відпрацьованих ртутьвмісних ламп (енергоощадних) як відходам I класу небезпеки, що збираються і обов’язково передаються на утилізацію підприємствам, що мають ліцензію на поводження з такими небезпечними відходами.

Всі відходи, що утворюються в процесі діяльності/роботи, підлягають обліку.

Вимоги безпеки при поводженні з відходами:

Під час роботи з відходами (прибирання виробничих приміщень, збір і сортування, навантаження, транспортування, розвантаження та ін.) працівники повинні бути забезпечені засобами індивідуального захисту та дотримуватися вимог інструкцій з охорони праці, що діють на підприємстві.

Наведено перелік деяких відходів, які передаються на утилізацію організаціям, які мають ліцензію на поводження з відходами як вторинної сировини:

- Макулатура;
- Матеріали пакувальні вторинні

Відвантаження таких відходів здійснюється відповідно до договору (контракту).

Побутові та будівельні відходи вивозяться на полігон твердих побутових відходів міста, також відповідно до договору з комунальним дорожньо-експлуатаційним управлінням.

Особи, винні в порушенні встановленого порядку поводження з відходами (порушення правил обліку відходів, самовільне складування і видалення відходів, передача відходів в інші підприємства/організації з порушенням встановлених правил), згідно законодавства несуть дисциплінарну, адміністративну або кримінальну відповідальність.

4.8.3 Визначення впливу та заходів щодо поводження з відходами ІТ галузі

З метою визначення та прогнозування впливу відходів на навколишнє середовище, своєчасного виявлення негативних наслідків, їх запобігання відповідно до Закону України «Про відходи» повинен здійснюватися моніторинг місць утворення, зберігання, і видалення відходів. Відомості про місце утворення та місце розташування відходів зазначаються на «План схемі місці розміщення відходів організації / виробництва» та наводяться у таблиці Г.2 в додатку Г, а Відомості про склад і властивості відходів, що утворюються, а також ступінь їх небезпечності для навколишнього природного середовища та здоров'я людини у табл. Г.3 додатку Г.

Висновки до розділу

В результаті проведеної роботи було зроблено аналіз умов праці, шкідливих та небезпечних чинників, з якими стикається робітник. Було визначено параметри і певні

характеристики приміщення для роботи над запропонованим проектом написаному в дипломній роботі, описано, які заходи потрібно зробити для того, щоб дане приміщення відповідало необхідним нормам і було комфортним і безпечним для робітника. Приведені рекомендації щодо організації робочого місця, а також важливу інформацію щодо пожежної та електробезпеки. Була наведені, розміри приміщення та наведено значення температури, вологості й рухливості повітря, необхідна кількість і потужність ламп та інші параметри, значення яких впливає на умови праці робітника, а також – наведені інструкції з охорони праці, техніки безпеки при роботі на комп’ютері.

А також визначені основні екологічні аспекти впливу на навколишнє природнє середовище та зазначені заходи щодо поводження з ними.

ВИСНОВКИ

Метою магістерської роботи було дослідження виявлення проблем при роботі з даними з соціальних мереж та підвищення точності емоційної класифікації.

У ході роботи були отримані результати, за якими можна зробити наступні висновки:

1. Актуальність роботи обумовлена тим, що соціальні мережі грають важливу роль у житті людини. Перед тим, як купити той чи інший товар, покупець читає багато коментарів та відгуків та робить висновок купляти йому цей товар чи шукати інший. Простота розміщення текстів у соціальних мережах є причиною зростання кількості інформації, котру людина вже не в змозі обробити за короткий проміжок часу.
2. Встановлена проблема незбалансованості даних, яка впливає на роботу класифікатора.
3. Розглянуті програмні інструменти для роботи з соціальними мережами та даними з них. Для роботи були обрані Community версія Microsoft Microsoft Visual Studio та Microsoft Azure Machine Learning Studio.
4. Розглянуті математичні моделі і методи вирішення задачі виявлення тональності у тексті. В якості моделі класифікатора був обраний метод максимальної ентропії.
5. Розглянута методологія побудови тренувальної та прогнозуючою моделей класифікатора. Зроблено висновок, що для великої кількості даних, краще використовувати готові засоби машинного навчання.
6. Розроблено додаток , який збирає необхідні дані з соціальної мережі Twitter.
7. Проаналізовані деякі методи боротьби з незбалансованими даними.
8. На основі міри обчислення ваг ознак/термів у векторно-просторовій моделі дельта TF-IDF розроблено додаток, який обчислює ці ваги. Необхідністю розробки такого додатку стало те, що ця міра не реалізована в сервісі, який використовується для побудови прогнозуючої моделі.
9. Виявлено, що розглянута міра обчислення ваг ознак/термів у векторно-просторовій моделі дельта TF-IDF є найефективнішою.

В якості продовження роботи можна й надалі покращувати точність класифікації, використовуючи інші класифікатори, нейронні мережі, інші способи векторизації тощо. Готову модель прогнозування можна буде використовувати через API, який надає Microsoft Azure Machine Learning Studio, що надасть можливість використовувати її як на різних веб-сервісах так і у десктопних додатках.

ПЕРЕЛІК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Пользовательский контент [Електронний ресурс] / Википедия - Режим доступу: www.URL: https://ru.wikipedia.org/wikiПользовательский_контент - 25.10.2017 р.

2. A Practical Approach for Content Mining of Tweets // Sunmoo Yoon, Suzanne Bakken // American journal of preventive medicine. - 2013. - Vol. 45. – Issue 1 (136). – P. 122-129.

3. Sentiment Analysis [Електронний ресурс] / Wikipedia - Режим доступу: www.URL: https://en.wikipedia.org/wiki/Sentiment_analysis - 30.10.2017 р.

4. Thumbs up?: sentiment classification using machine learning techniques / Bo Pang, Lilian & Shivakumar Vaithyanathan // Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10. Association for Computational Linguistics. – 2002. – P.79-86

5. Задача классификации [Електронний ресурс] / Википедия - Режим доступу: www.URL: https://ru.wikipedia.org/wiki/Задача_классификации - 24.11.2017 р.

6. Sentistrenth documentation [Електронний ресурс] / Sentistrenth - Режим доступу: www.URL: http://sentistrength.wlv.ac.uk/documentation/ - 24.11.2017 р.

7. 8 Tactics to Combat Imbalanced Classes in Your Machine Learning Dataset [Електронний ресурс] / Machinelearningmastery - Режим доступу: www.URL: https://machinelearningmastery.com/tactics-to-combat-imbalanced-classes-in-your-machine-learning-dataset/ - 27.11.2017 р.

8. API [Електронний ресурс] / Wikipedia - Режим доступу: www.URL: https://en.wikipedia.org/wiki/Application_programming_interface - 17.11.2017 р.

9. Twitter documentation [Електронний ресурс] / Twitter - Режим доступу: www.URL: https://developer.twitter.com/en/docs - 17.11.2017 р.

10. Twitter Firehose vs Twitter API [Електронний ресурс] / Brigplanet - Режим доступу: www.URL: https://brightplanet.com/2013/06/twitter-firehose-vs-twitter-api-whats-the-difference-and-why-should-you-care/ - 19.11.2017 р.

11. Discovertext [Електронний ресурс] / Discovertext - Режим доступу: www.URL: https://discovertext.com/pricing/ - 19.11.2017 р.

12. About discovertext gnip [Електронний ресурс] / Discover Text - Режим доступу: www.URL: https://discovertext.com/gnip-enabled-access-for-discovertext-users/ - 20.11.2017 р.

13. Visual Studio [Електронний ресурс] / Visual studio - Режим доступу: www.URL: https://www.visualstudio.com/ru/vs/pricing/ - 20.11.2017 р.

14. Microsoft Azure [Електронний ресурс] / Microsoft - Режим доступу: www.URL: https://azure.microsoft.com/en-us/pricing/details/machine-learning-studio/ - 21.11.2017 р.

15. Azure ML: A brief Introduction [Електронний ресурс] / Project Botticelli - Режим доступу: www.URL: https://projectbotticelli.com/knowledge/brief-introduction-to-microsoft-azure-ml?pb_campaign=pb2014vyru - 22.11.2017 р.

16. Determining Word–Emotion Associations from Tweets by Multi-Label Classification / Felipe Bravo-Marquez, Eibe Frank, Saif M. Mohammad, Bernhard Pfahringer // IEEE/WIC/ACM International Conference on Web Intelligence. – 13-16 Oct. 2016. – P536-539.

17. Векторна модель [Електронний ресурс] / Вікіпедія - Режим доступу: www.URL: https://uk.wikipedia.org/wiki/Векторна_модель - 24.11.2017 р.

18. Современные методы анализа тональности текста [Електронний ресурс] / Data Review - Режим доступу: www.URL: http://datareview.info/article/sovremennyie-metodyi-analiza-tonalnosti-teksta/ - 24.11.2017 р.

19. Немного про word2vec: полезная теория [Електронний ресурс] / Nlpx - Режим доступу: www.URL: http://nlpx.net/archives/179 - 25.11.2017 р.

20. Brown Clustering [Електронний ресурс] / Wikipedia - Режим доступу: www.URL: https://en.wikipedia.org/wiki/Brown_clustering - 26.11.2017 р.

21. Twitter Sentiment Classification using Distant Supervision / Alex Go, Richa Bhayani, Lei Huang - January 2009

22. Delta TFIDF: An Improved Feature Space for Sentiment Analysis / Justin Martineau, Tim Finin // Proceedings of the Third AAAI International Conference on Weblogs and Social Media – 17 May, 2009

23. Machine Learning Tutorial: The Naïve Bayes Text Classifier [Електронний ресурс] / Datumbox blog - Режим доступу: www.URL: http://blog.datumbox.com/machine-learning-tutorial-the-naive-bayes-text-classifier/ - 27.11.2017 р.

24. Sentiment Symposium Tutorial: Classifiers [Електронний ресурс] / Christopher Potts - Режим доступу: www.URL: http://sentiment.christopherpotts.net/classifiers.html - 27.11.2017 р.

25. Azure documentation [Електронний ресурс] / Microsoft - Режим доступу: www.URL: https://msdn.microsoft.com/en-us/library/azure/dn905994.aspx?f=255&MSPPError=-2147217396 - 22.11.2017 р.

26. Logistic Regression for Machine Learning [Електронний ресурс] / Machine Learning Mastery - Режим доступу: www.URL: https://machinelearningmastery.com/logistic-regression-for-machine-learning/ - 28.11.2017 р.

27. Speech and Language Processing / Daniel Jurafsky, James H. Martin. - Stanford University, University of Colorado at Boulder, 2017

28. Шумові слова [Електронний ресурс] / Вікіпедія - Режим доступу: www.URL: https://uk.wikipedia.org/wiki/Шумові_слова - 29.11.2017 р.

29. Регулярные выражения [Електронний ресурс] / Википедия - Режим доступу: www.URL: https://ru.wikipedia.org/wiki/Регулярные_выражения - 29.11.2017 р.

30. Морфологія [Електронний ресурс] / Вікіпедія - Режим доступу: www.URL: https://uk.wikipedia.org/wiki/Морфологія - 30.11.2017 р.

31. An algorithm for suffix stripping. Program / Porter, M. F., 1980., 14(3), 130–127

32. Machine Learning : Text feature extraction (tf-idf) - Part II [Електронний ресурс] / Christian Perone blog - Режим доступу: www.URL: http://blog.christianperone.com/2011/10/machine-learning-text-feature-extraction-tf-idf-part-ii/ - 30.11.2017 р.

33. Chi Square test for feature selection [Електронний ресурс] / Learn4master - Режим доступу: www.URL: http://www.learn4master.com/machine-learning/chi-square-test-for-feature-selection - 01.12.2017 р.

34. The chi-square test [Електронний ресурс] / Stanford - Режим доступу: www.URL: https://web.stanford.edu/class/psych252/cheatsheets/chisquare.html - 07.12.2017 р.

35. Simple guide to confusion matrix terminology [Електронний ресурс] / Dataschool - Режим доступу: www.URL: http://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/ - 09.12.2017 р.

36. Оценка классификатора (точность, полнота, F-мера) [Електронний ресурс] / Bazhenov blog - Режим доступу: www.URL: http://bazhenov.me/blog/2012/07/21/classification-performance-evaluation.html - 10.12.2017 р.

37. A systematic analysis of performance measures for classification tasks / Marina Sokolova, Guy Lapalme // Information Processing and Management 45 (2009) 427–437

38. Micro- and Macro-average of Precision, Recall and F-Score [Електронний ресурс] / Rushdishams blog - Режим доступу: www.URL: http://rushdishams.blogspot.in/2011/08/micro-and-macro-average-of-precision.html - 12.12.2017 р.

39. scikit-learn [Електронний ресурс] / Scikit-learn - Режим доступу: www.URL: http://scikit-learn.org/stable/ - 13.12.2017 р.

40. DeltaTfidfVectorizer for scikit-learn [Електронний ресурс] / Python scrypts - Режим доступу: www.URL: https://pypi.python.org/pypi/sklearn-deltatfidf/0.1 - 13.12.2017 р.

41. Python [Електронний ресурс] / Вікіпедія - Режим доступу: www.URL: https://uk.wikipedia.org/wiki/Python - 15.12.2017 р.

42. scikit-learn [Електронний ресурс] / Wikipedia - Режим доступу: www.URL: https://en.wikipedia.org/wiki/Scikit-learn - 15.12.2017 р.

43. openpyxl Documentation [Електронний ресурс] / Openpyxl - Режим доступу: www.URL: https://media.readthedocs.org/pdf/openpyxl/latest/openpyxl.pdf - 17.12.2017 р.

44. Euclidean norm learn [Електронний ресурс] / Wikipedia - Режим доступу: www.URL: https://en.wikipedia.org/wiki/Norm_%28mathematics%29#Euclidean_norm - 18.12.2017 р.

45. Pycharm [Електронний ресурс] / Вікіпедія - Режим доступу: www.URL: https://uk.wikipedia.org/wiki/PyCharm - 18.12.2017 р.

46. Типове положення про порядок проведення навчання і перевірки знань з питань охорони праці (НПАОП 0.00-4.12-05) [Електронний ресурс] / Законодавство України - Режим доступу: www.URL: http://zakon0.rada.gov.ua/laws/show/z0231-05 - 21.12.2017 р.

47. Типове положення про інструктажі, спеціальне навчання та перевірку знань з питань пожежної безпеки на підприємствах, в установах та організаціях України (НАПБ Б.02.005-2003) [Електронний ресурс] / Законодавство України - Режим доступу: www.URL: http://zakon0.rada.gov.ua/laws/show/z1148-03 - 21.12.2017 р.

48. Санітарні норми мікроклімату виробничих приміщень (ДСН 3.3.6.042.-99) [Електронний ресурс] / Законы Украины - Режим доступу: www.URL: http://uazakon.com/documents/date_42/pg_ikcfxj.htm - 22.12.2017 р.

49. Правила і норми роботи з візуальними дисплейними терміналами електронно-обчислювальних машин (ДСанПіН 3.3.2.007-98) [Електронний ресурс] / Педрада - Режим доступу: www.URL: http://zakon.pedrada.com.ua/regulations/10637/478672/ - 22.12.2017 р.

50. Норми визначення категорій приміщень, будинків та зовнішніх установок за вибухопожежною та пожежною небезпекою (НАПБ Б.03.002-2007) [Електронний ресурс] / ДНАОП - Режим доступу: www.URL: https://dnaop.com/html/32980/doc-НАПБ_Б.03.002.-2007 - 23.12.2017 р.

51. Санітарні норми мікроклімату виробничих приміщень (ДСН 3.3.6.042-99) [Електронний ресурс] / UAinfo - Режим доступу: www.URL: http://ua-info.biz/legal/basetp/ua-zmptae.htm - 23.12.2017 р.

52. Санітарні норми виробничого шуму, ультразвуку та інфразвуку (ДСН 3.3.6.037-99) [Електронний ресурс] / Нормативно-директивні документи МОЗ України - Режим доступу: www.URL: http://mozdocs.kiev.ua/view.php?id=1789 - 23.12.2017 р.

53. Про затвердження Положення про розробку інструкцій з охорони праці (ДНАОП 0.00-4.15-98) [Електронний ресурс] / Законодавство України – Режим доступу: www.URL: http://zakon2.rada.gov.ua/laws/show/z0226-98 - 23.12.2017 р.

54. Про затвердження Правил охорони праці під час експлуатації електронно-обчислювальних машин (НПАОП 0.00-1.28-10) [Електронний ресурс] / Законодавство України – Режим доступу: www.URL: http://zakon2.rada.gov.ua/laws/show/z0293-10 - 23.12.2017 р.

55. Природне і штучне освітлення (ДБН В.2.5-28:2015) [Електронний ресурс] / Державні будівельні норми України – Режим доступу: www.URL: http://dbn.at.ua/load/normativy/dbn/dbn_v_2_5_28_2015/1-1-0-1188 - 23.12.2017 р.

56. Державні санітарні правила і норми (ДСанПіН 2.2.7.029) [Електронний ресурс] / LIGA:ZAKON – Режим доступу: www.URL: http://search.ligazakon.ua/l_doc2.nsf/link1/MOZ4153.html - 25.12.2017 р.

57. Державні санітарні норми виробничої загальної та локальної вібрації (ДСН 3.3.6.039-99) [Електронний ресурс] / ДНАОП - Режим доступу: www.URL: https://dnaop.com/html/31680/doc-%D0%94%D0%A1%D0%9D_3.3.6.039-99 - 25.12.2017 р.

58. ССБТ. Электромагнитные поля радиочастот. Допустимые уровни на рабочих местах и требования к проведению контроля (ГОСТ 12.1.006-84) [Електронний ресурс] / ДНАОП - Режим доступу: www.URL: https://dnaop.com/html/42235/doc-%D0%93%D0%9E%D0%A1%D0%A2_12.1.006-84 - 26.12.2017 р.

59. Электробезопасность. Защитное заземление. Зануление (ГОСТ 12.1.030-81) [Електронний ресурс] / ДНАОП - Режим доступу: www.URL: https://dnaop.com/html/2128/doc-%D0%93%D0%9E%D0%A1%D0%A2_12.1.030-81 - 26.12.2017 р.

60. Нормы качества электрической энергии в системах электроснабжения общего назначения (ГОСТ 13109-97) [Електронний ресурс] / ДНАОП - Режим доступу: www.URL: https://dnaop.com/html/42313/doc-%D0%93%D0%9E%D0%A1%D0%A2_13109-97 - 26.12.2017 р.

61. Общие санитарно-гигиенические требования к воздуху рабочей зоны (ГОСТ 12.1.005-88) [Електронний ресурс] / document.UA - Режим доступу: www.URL: http://document.ua/ssbt_-obshie-sanitarno-gigienicheskie-trebovanija-k-vozduhu--nor3205.html - 26.12.2017 р.

62. Опалення, вентиляція та кондиціювання (ДБН В.2.5-67:2013) [Електронний ресурс] / ДНАОП - Режим доступу: www.URL: https://dnaop.com/html/32609/doc-%D0%94%D0%91%D0%9D_%D0%92.2.5-67_2013 - 27.12.2017 р.

63. Правила безпечної експлуатації електроустановок споживачів (НПАОП 40.1-1.21-98) [Електронний ресурс] / ДНАОП - Режим доступу: www.URL:

https://dnaop.com/html/2029/doc-%D0%9D%D0%9F%D0%90%D0%9E%D0%9F_40.1-1.21-98 - 27.12.2017 р.

64. ССБТ. Пожаровзырвоопасность веществ и материалов (ГОСТ 12.1.044-89) [Електронний ресурс] / STROYNOTE строительный портал - Режим доступу: www.URL: http://www.stroynote.com.ua/construction-regulations/document-1611.html - 27.12.2017 р.

65. Вредные химические вещества. Неорганические соединения элементов / I-IV групп: Справ. изд. / Под ред. В.А.Филова и др.-Л: Химия, 1986. -512с.

66. Свойства неорганических соединений элементов. Справочник/ Ефимов А.И. и др.-Л: Химия, 1983.-392с.

ДОДАТОК А

Блок-схеми алгоритмів

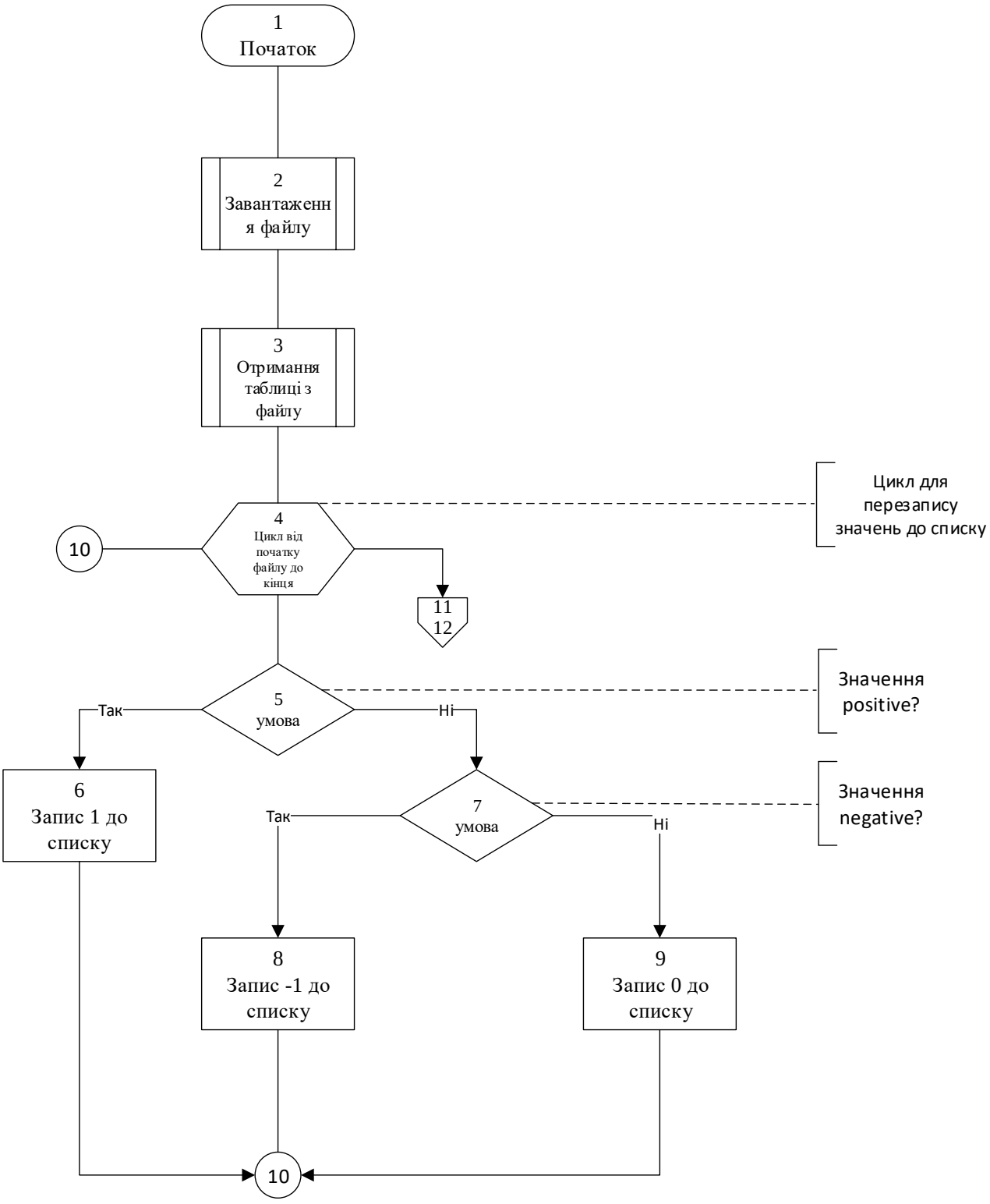

Рисунок А1 – Блок-схема алгоритму створення даних для тренування моделі

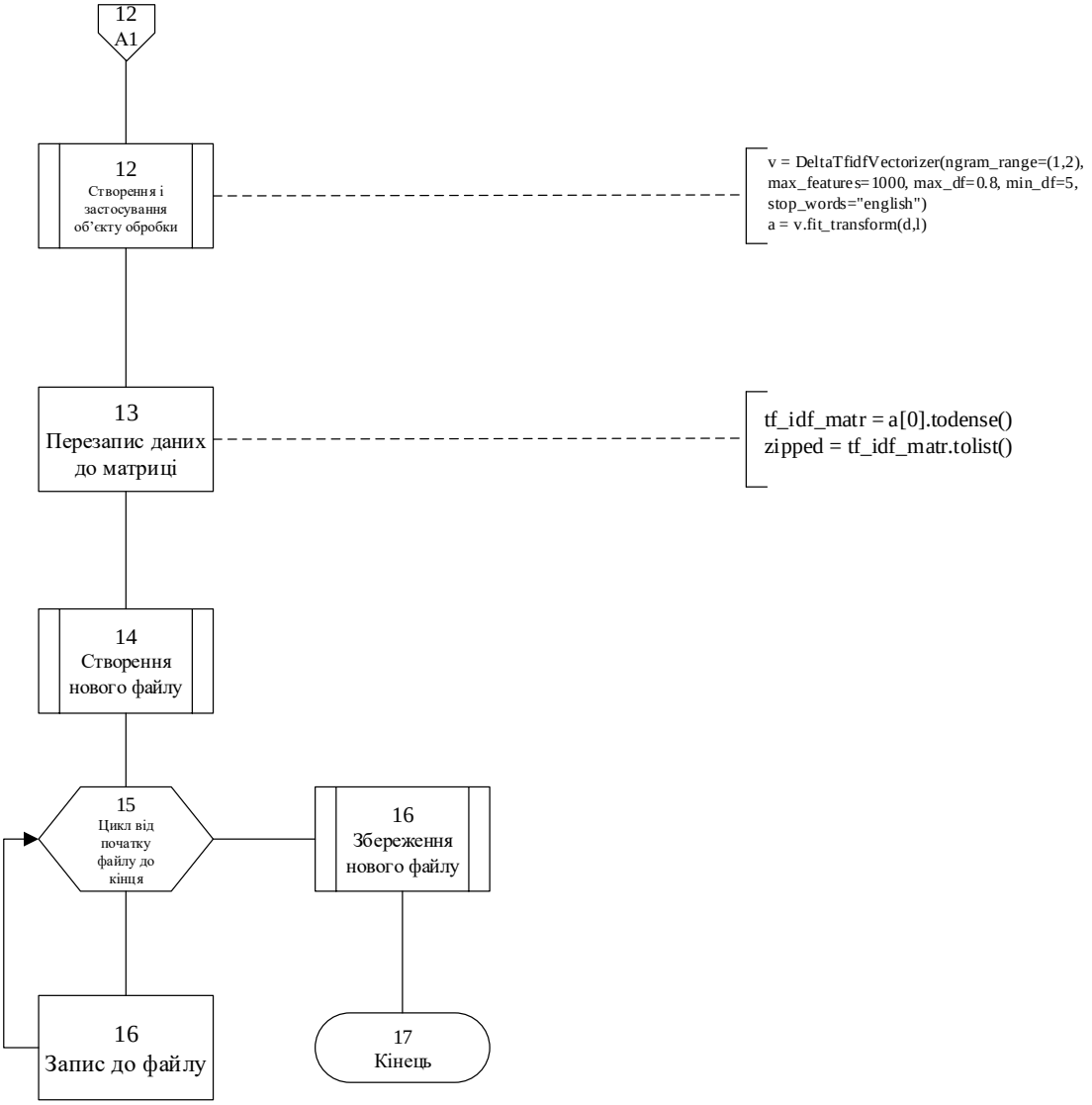

Рисунок А2 – Продовження блок-схеми алгоритму створення даних для тренування моделі

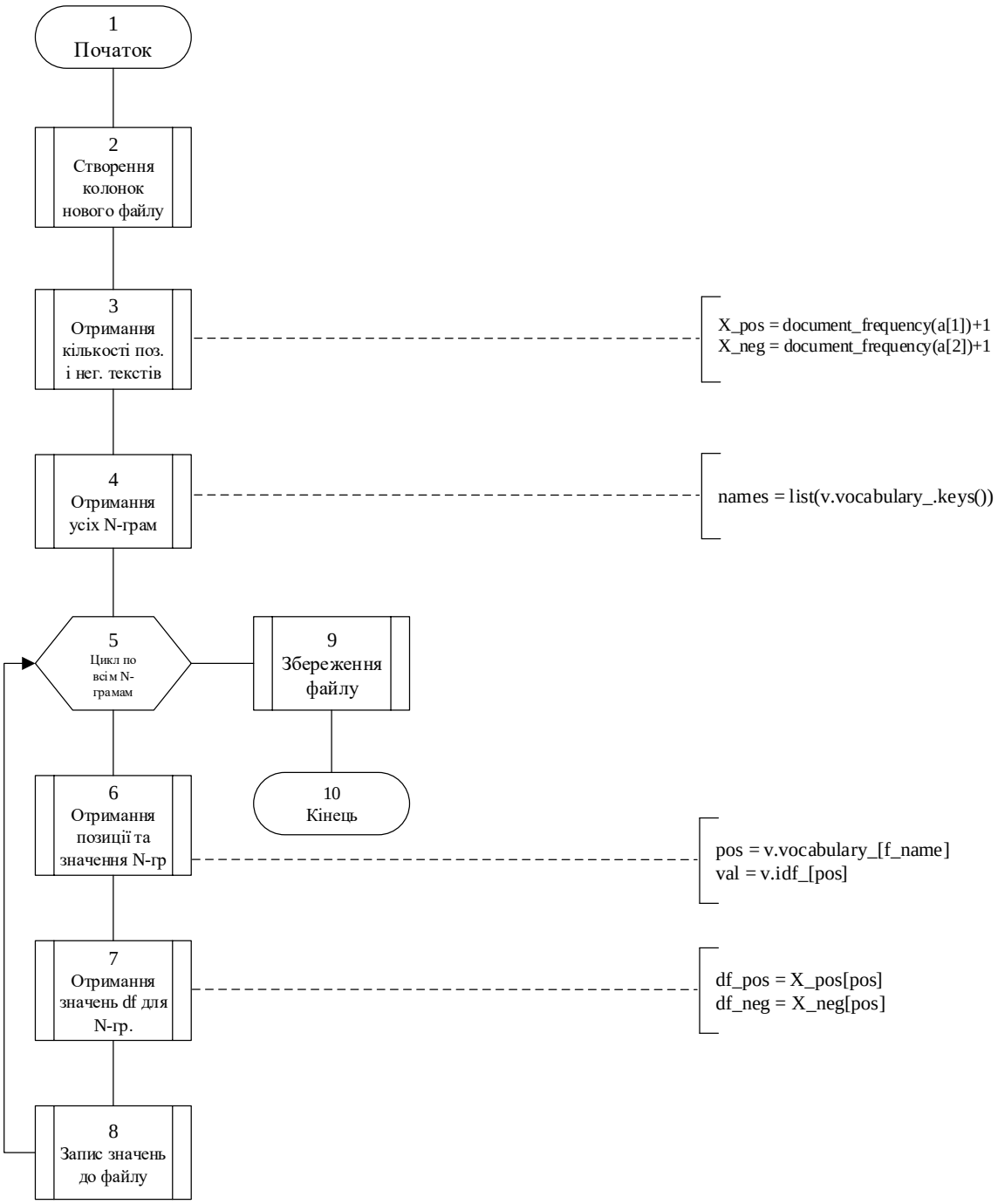

Рисунок А3 – Блок-схема алгоритму створення словника у вигляді файлу

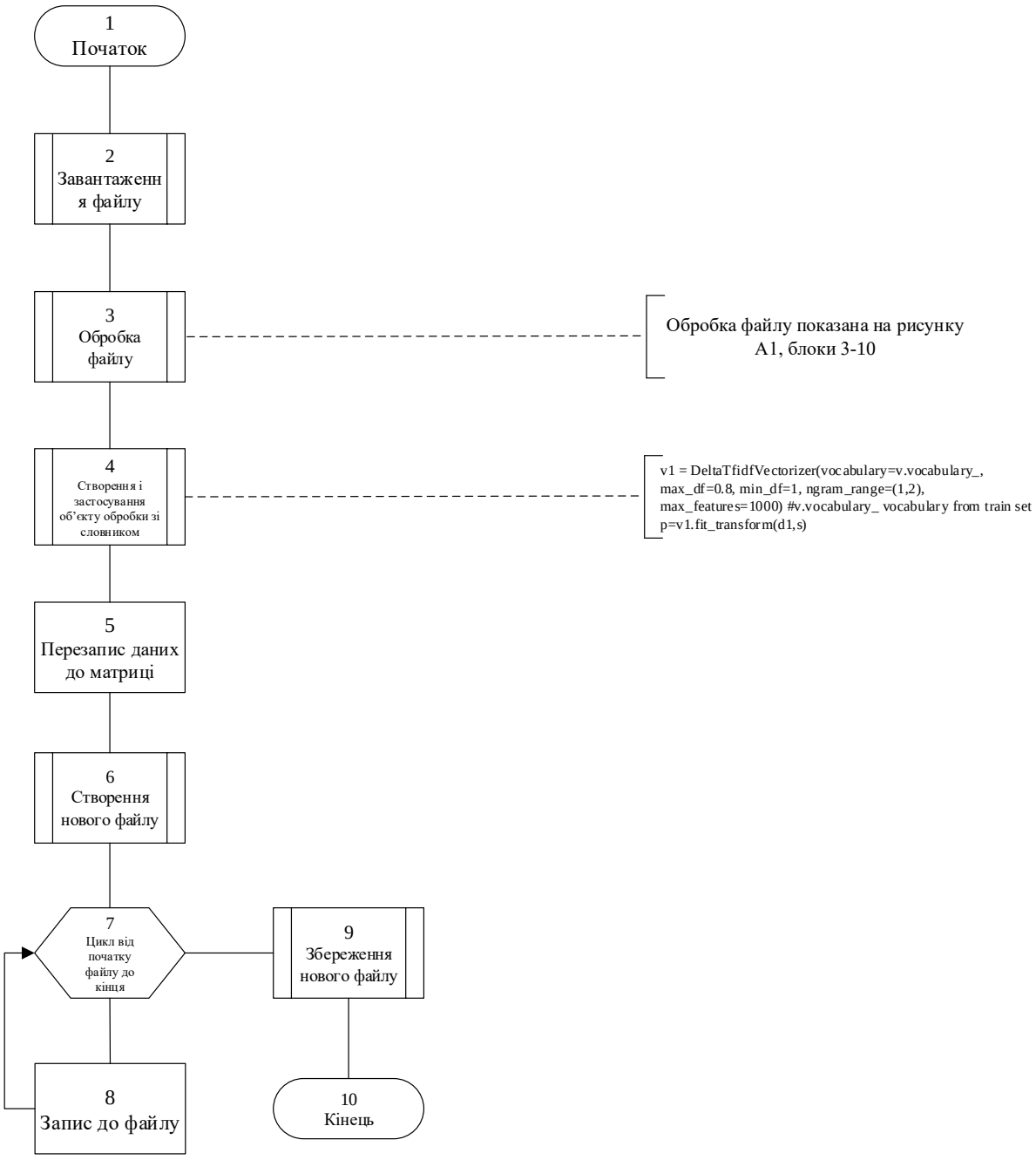

Рисунок А4 – Блок-схема алгоритму створення даних для оцінки моделі

ДОДАТОК Б

Лістинг методу отримання даних з Twitter та функції main

```
static async void getStreamOfTweets()
        {
            TwitContext db = new TwitContext();

            var auth = new SingleUserAuthorizer
            {
               CredentialStore = new SingleUserInMemoryCredentialStore
               {
                  ConsumerKey = "key",
                  ConsumerSecret = "secret",
                  AccessToken = "token",
                  AccessTokenSecret = "tokensecret"
               }
            };
            var context = new TwitterContext(auth);
            int count = 0;

            var st = await
                (from strm in context.Streaming
                 where strm.Type == StreamingType.Sample && //(StreamingType.Filter поиск
по ключевым словам)
                 strm.Language == "en"
                 select strm)
                .StartAsync(async strm =>
                {
                   JObject obj = JObject.Parse(strm.Content);
                   bool isRetweeted = obj["text"].ToString().Contains("RT");
                  Console.WriteLine("User id: " + obj["user"]["id"] + "\n" + "User name: " +
obj["user"]["name"] + "\n" + "User screen name: " + obj["user"]["screen_name"] + "\n"
                      + "Post id: " + obj["id"] + "\n" + "Is retweeted: " + isRetweeted.ToString()
+ "\n"
                      + "Date: " + obj["created_at"] + "\n" + "Text: " + obj["text"] + "\n" +
"Source: " + "https://twitter.com/" + obj["user"]["screen_name"] + "/status/" + obj["id"] +
"\n");
                   Console.WriteLine(db.Tweets.Count());
                   HashTagEntity b = new HashTagEntity();
                   var sd = b.Tag;
                   UserTwt user = new UserTwt()
                  {
                     UserId = obj["user"]["id"].ToString(),
                     UserName = obj["user"]["name"].ToString(),
                     UserScreenName = obj["user"]["screen_name"].ToString()
                  };
                   TweetTxt tweet = new TweetTxt()
                   {
                      PostId = obj["id"].ToString(),
                      IsRetwited = isRetweeted,
                      Date = obj["created_at"].ToString(),
```

```
                    Text = obj["text"].ToString(),
                    Source = "https://twitter.com/" + obj["user"]["screen_name"] + "/status/" +
obj["id"],
                    UoloadDate = DateTime.Now,
                    UsTwt = user
                  };
                  try
                {
                    db.Users.Add(user);
                    db.Tweets.Add(tweet);
                    db.SaveChanges();
                }
                catch (DbEntityValidationException e)
                {
                    Console.Write(e.EntityValidationErrors.First().ValidationErrors);
                }
                  DbClear c = new DbClear();
                  c.ClearRT(tweet);
                  c.ClearSource(tweet);
                  c.ClearEmoticos(tweet);
              });

          }

        static void Main(string[] args)
        {
            TwitContext db = new TwitContext();
            Console.WriteLine(db.Tweets.Count());
            var auth = new SingleUserAuthorizer
            {
                CredentialStore = new SingleUserInMemoryCredentialStore
                {
                    ConsumerKey = "OTeXEesPz4Xv6LlotPaaIcmdv",
                    ConsumerSecret =
"XpwkpjovRFKvUQyqTtFy1jALLQDHP0WOfCdIfaVdmengBuQAjy",
                    AccessToken = "878966673621102592-
oTpEO7yRtUTCghJKc7uaWz601fxDjnb",
                    AccessTokenSecret =
"brweXGq4mwB1XbOc1oHpNsH6ZV5KSecHARneDBITEVEnc"
                }
            };

            Task task = new Task(getStreamOfTweets);
            task.Start();
            task.Wait();
            Console.ReadLine();

```

ДОДАТОК В

Лістинг скрипту для розрахунку дельта TF-IDF

```
from sklearn_deltatfidf import DeltaTfidfVectorizer
import numpy as np
import scipy.sparse as sp
from openpyxl import load_workbook
import openpyxl

def document_frequency(X):
    """Count the number of non-zero values for each feature in sparse X."""
    if sp.isspmatrix_csr(X):
        return np.bincount(X.indices, minlength=X.shape[1])
    else:
        return np.diff(sp.csc_matrix(X, copy=False).indptr)

wb=load_workbook('C:/Users/Admin/Desktop/data_preprocessed (delta tfidf) train.xlsx')
wb1=load_workbook('C:/Users/Admin/Desktop/data_preprocessed (delta tfidf) score.xlsx')
wb2=load_workbook('C:/Users/Admin/Desktop/NewData_preprocessed.xlsx')
sheet=wb.get_sheet_by_name("data")
sheet_s=wb1.get_sheet_by_name("Worksheet")
sheet_s1=wb2.get_sheet_by_name("Worksheet")

v = DeltaTfidfVectorizer(ngram_range=(1,2), max_features=1000, max_df=0.8, min_df=5,
stop_words="english")

d = list()
d1 = list()
s=list()
l = list()

for i in range(2,5894): #5892 texts from data, 2 because row nuber starts from 1, first row is
names of cols
    d.append(sheet.cell(row = i, column = 5).value)
    if(sheet.cell(row = i, column = 4).value=="positive"):
        l.append(1)
    elif sheet.cell(row = i, column = 4).value=="negative":
            l.append(-1)
    else:
         if(sheet.cell(row=i, column=4).value == "neutral"):
             l.append(0)
for i in range(2,5898): #5897 texts from data, 2 because row nuber starts from 1, first row is
names of cols
    d1.append(sheet_s.cell(row = i, column = 5).value)
    if (sheet_s.cell(row=i, column=4).value == "positive"):
        s.append(1)
    elif sheet_s.cell(row=i, column=4).value == "negative":
        s.append(-1)
    else:
         if (sheet_s.cell(row=i, column=4).value == "neutral"):
            s.append(0)
```

```
l.append(1)
l.append(0)
l.append(1)
l.append(0)

wb_new = openpyxl.Workbook()
wb_new1 = openpyxl.Workbook()
wb_new2 = openpyxl.Workbook()
wb_new3 = openpyxl.Workbook()
sheet = wb_new.create_sheet("data")
sheet_1 = wb_new1.create_sheet("data")
sheet_2 = wb_new2.create_sheet("data")
sheet_3 = wb_new3.create_sheet("data")

#check elements
print(d[0])
print(d[5891])
print(l[5891])

#Delta TF_idf for train data
a = v.fit_transform(d,l)

tf_idf_matr = a[0].todense()
zipped = tf_idf_matr.tolist()
matr_size = tf_idf_matr.shape[0]

c1=1
for i in range(0,5892):
    for j in range(0,999):
        sheet.cell(row=c1, column=j+1).value = zipped[i][j]
    c1=c1+1
wb_new.save('C:/Users/Admin/Desktop/data_new.xlsx')

#Getting feature names and create vocabulary
sheet_2.cell(row=1, column=1).value = 'id'
sheet_2.cell(row=1, column=2).value = 'Ngram'
sheet_2.cell(row=1, column=3).value = 'DF'
names = list(v.vocabulary_.keys())
df = 0;
new_matr=list()
new_matr = zipped.copy()

for i, f_name in enumerate(names):
    pos = v.vocabulary_[f_name]
    val = v.idf_[pos]
    for j in range(0,len(zipped)):
        if zipped[j][pos] != 0:
            sheet_2.cell(row=i + 2, column=4).value = zipped[j][pos]
            print(j,pos,'--',zipped[j][pos])
    sheet_2.cell(row=i+2, column=1).value = i+1
    sheet_2.cell(row=i+2, column=2).value = f_name
    df=0
```

```
idf = v.idf_.tolist()
for i in range(0,999):
    df = idf.count(idf[i])
    sheet_2.cell(row=i + 2, column=3).value = df
wb_new2.save('C:/Users/Admin/Desktop/vocabulary_score_1.xlsx')

sheet_1.cell(row=1, column=1).value = 'id'
sheet_1.cell(row=1, column=2).value = 'Ngram'
sheet_1.cell(row=1, column=3).value = 'DF_pos'
sheet_1.cell(row=1, column=4).value = 'DF_neg'
sheet_1.cell(row=1, column=5).value = 'IDF'
X_pos = document_frequency(a[1])+1
X_neg = document_frequency(a[2])+1
for i, f_name in enumerate(names):
    pos = v.vocabulary_[f_name]
    val = v.idf_[pos]
    df_pos = X_pos[pos]
    df_neg = X_neg[pos]
    sheet_1.cell(row=i+2, column=1).value = i+1
    sheet_1.cell(row=i+2, column=2).value = f_name
    sheet_1.cell(row=i + 2, column=3).value = df_pos
    sheet_1.cell(row=i + 2, column=4).value = df_neg
    sheet_1.cell(row=i+2, column=5).value = val
wb_new1.save('C:/Users/Admin/Desktop/vocabulary_score_4(df pos neg).xlsx')

#Features with scores for score data
v1 = DeltaTfidfVectorizer(vocabulary=v.vocabulary_, max_df=0.8, min_df=1,
ngram_range=(1,2), max_features=1000) #v.vocabulary_ vocabulary from train set
p=v1.fit_transform(d1,s)
print(v1.vocabulary_)
print(v1.idf_)
tf_idf_matr1 = p[0].todense()
zipped1 = tf_idf_matr1.tolist()
print(tf_idf_matr1.shape)
c1=1
for i in range(0,5896):
    for j in range(0,999):
        sheet.cell(row=c1, column=j+1).value = zipped1[i][j]
    c1=c1+1
wb_new.save('C:/Users/Admin/Desktop/score_dataset_deltatfidf.xlsx')

#Getting feature names
names = list(v1.get_feature_names())
for i, f_name in enumerate(names):
    name = 'Preprocessed text.['+f_name+']'
    sheet_3.cell(row=1, column=i+1).value = 'Preprocessed text.['+f_name+']'
wb_new2.save('C:/Users/Admin/Desktop/names_for_score.xlsx')
```

ДОДАТОК Г

Таблиця Г.1 – Аналіз небезпечних і шкідливих виробничих факторів

Небезпечні і шкідливі виробничі фактори	Джерела факторів (види робіт)	Кількісна оцінка	Нормативні документи
1	2	3	4
фізичні			
- підвищена температура поверхонь обладнання	експлуатація ЕОМ, принтерів, сканерів чи/або серверного обладнання для роботи	2	[51]
- підвищений рівень шуму на робочому місці	-//-	2	[52]
- підвищений рівень вібрації	-//-	2	[57]
- підвищена або знижена вологість повітря	-//-	2	[51]
- підвищена або знижена рухливість повітря	-//-	1	[51]
- підвищений рівень іонізуючого випромінення в робочій зоні	-//-	2	[51] [58]
- підвищений рівень електромагнітного випромінення	-//-	2	[58]
- підвищений рівень напруги електричної мережі, замикання якої може відбутися через тіло людини	-//-	4	[59] [60]
- підвищена напруженість електричного поля	-//-	2	[58]
- підвищена напруженість магнітного поля	-//-	2	[58]
- недостатність природного світла	порушення умов праці (вимог до приміщень)	2	[55]
- недостатнє освітлення робочої зони	порушення гігієнічних параметрів виробничого середовища	3	[55]

Продовження таблиці Г.1

Небезпечні і шкідливі виробничі фактори	Джерела факторів (види робіт)	Кількісна оцінка	Нормативні документи
1	2	3	4
- підвищена яскравість світла	порушення умов праці (організації місця праці- налагодження моніторів)	1	[49]
- понижена контрастність	-//-	1	[49]
хімічні:			
- загазованість повітря робочої зони, яка впливає на організм людини через органи дихання та надає токсичну і канцерогенну дію	від експлуатації сканерів, принтерів для роботи – O_3, оплавлення електричних і комутаційних кабелів, резисторів, конденсаторів, напівпровідникових діодів, транзисторів й інше в ЕОМ та системах кондиціонування повітря - CO, CO_2, SO_2, P_2O_5, H_2S, HCl, H, NH_3, ClF_3, F_2O_2, F_2O_3, SeO_2. SeF_6, TeF_6, $COCl_2$, SO_2F_2, інш.	3	[63] [62] [61] [64]
психофізіологічні:			
- нервово-психічна перевантаження (розумове, перенапруження аналізаторів-зорових)	- пошук інформації для постановки теми; - пошук та аналіз аналогів і літератури; - пошук наявних технологій, моделювання та аналіз алгоритмів; - виконання роботи за темою диплома, тестування; - оформлення роботи	4	[54] [49]
- фізичні (статичне – сидіння)	порушення умов праці (організації місця праці- сидіння користувача,) та організації робочого часу - безпервна робота)	2	[54] [49]

Таблиця Г.2 - Відомості про місце утворення та місце розташування відходів

№ з/п	Код та найменування відходів за ДК -005-96	Технологічний процес або виробництво, де утворюються відходи/клас небезпеки	Місце розташування відходу, тара та її кількість, місткість, розміри у разі наявності майданчиків розташування відходів необхідно зазначити тип покриття та наявність даху)	№ на схемі (додається масштабна схема місць розміщення відходів)
1	2	3	5	7
1	7710.3.1.26 Лампи люмінесцентні, та відходи, які містять ртуть, інші зіпсовані або відпрацьовані (Відпрацьовані ртутьвмісні люмінесцентні лампи)	1	буд.84, в приміщені кладової S=100м^2	8401-ТХ
2	7720.3.1.01 Відходи комунальні (міські) змішані, у т.ч. сміття з урн (Побутові відходи)	4	зовнішній майданчик зберігання побутових відходів біля буд .84 S=12м^2 V= 2,08м^3- 5 од.	8401-ТХ
3	7710.3.1.01 Макулатура паперова та картонна (Макулатура)		буд .84 4 поверх кім. 412 S =5,0 м.^{2}	8401-ТХ
4	7730.3.1.02 Матеріали пакувальні пластмасові зіпсовані, відпрацьовані чи забруднені (Матеріали пакувальні забруднені)	4	буд.84, контейнер V=0,9м^3 (1 од.)	8401-ТХ

Продовження таблиці Г.2

№ з/п	Код та найменування відходів за ДК -005-96	Технологічний процес або виробництво, де утворюються відходи/клас небезпеки	Місце розташування відходу, тара та її кількість, місткість, розміри у разі наявності майданчиків розташування відходів необхідно зазначити тип покриття та наявність даху)	№ на схемі (додається масштабна схема місць розміщення відходів)
1	2	3	5	7
5	Пакувальні матеріали батарейки Відходи друкуючих пристроїв.	4	буд. 84, кім. 412 m=5,0 кг.	8401-ТХ
6	Пакувальні матеріали, що не вміщують целюлозу	4	буд. 84, кім. 412 S =5,0 м.^{2}	8401-ТХ
7	Батарейки та акумулятори (малі)	3	буд. 84, кім. 412 V=0,0005 м^3	8401-ТХ
8	Відходи друкуючих пристроїв.	4	буд. 84, кім. 412 V=1,0 м^3	8401-ТХ

Таблиця Г.3 – Відомості про склад і властивості відходів, що утворюються, а також ступінь їх небезпечності для навколишнього природного середовища та здоров'я людини

№ п/п 148	Назва відходу	Клас небезпеки	Хімічний (у долях відсотків складників або інших одиницях виміру) та морфологічний склад	Фізико-хімічні властивості	Негативний вплив на навколишнє середовище та здоров'я людини
	2		4	5	6
	Відпрацьовані люмінесцентні лампи	I	**Ртуть** - 0,013 **Hg** **Скло** - 98,787 **(Na, K)$_2$ O 2SiO$_2$** **Алюміній** - 1,2 **Al**	**Ртуть** - $T_{кип.}$ = 356,58°C $T_{плав.}$ = **- 38,87°C** **Скло** - $T_{плав.}$ = 800°C **Алюміній** - $T_{кип.}$ = 2348°C $T_{плав.}$ = 660,1°C	Негативний вплив на ОС і людини визначається його хімічним складом. **Ртуть** У природних водах міститься в концентрації 0,00003 ... 0,0028 мг / л. Являючись потужним кумулятивним отрутою, з можливою канцерогенною і мутагенною дією. Процеси самоочищення водойм порушують концентрація ртуті понад 0,018 мг / л, порогова концентрація ртуті за впливом на санітарний режим водойм-0,01 мг / л. Наприкінці концентрація понад 0,03 є токсичною практично для всіх видів водних організмів. Надзвичайно токсична при попаданні з питною водою для тепло-кровних організмів, надходження ртуті з питною водою в кількості 75,0 ... 300,0 мг / сут є смертельним. Відрізняється високою токсичністю для будь-яких форм життя. При отруєнні па-рами спостерігається слабкість, головний біль, біль в шлунку, роздратування по-чек, навіть нефрит; катаральні явища. Розвивається тремтіння рук, ніг, всього тіла. Виникає стан підвищеної психічної збудливості [65]. Пари ртуті проявляють нейротоксичність, особливо страждають вищі відділи нервової системи. **Скло** Нетоксичні, безпечно в навколишньому середовищу, не шкідлива в нирках і водоймах. Вдихання скляного пилу (волокон) призводить до силікоз в зв'язку з високим вмістом сполук кремнію. Шкідливої дії не робить, але є небезпека механічних пошкоджень (порізи,

					травми). **Алюміній** Токсичний для водної біоти, теплокровних тварин і людей, в концентрації> 1 мг / л чинить негативний впливав на зростання с /г культур. У концентрації> 1 мг / л гальмує зростання мікрофлори водойм і стримує процеси самоочищення водойм. Рівень токсичності визначається формою, в якій знаходиться елемент. Впливає на обмін речовин і функції нервової системи. При попаданні на ґрунт, в воду і атмосферними повітря надає негативного впливу на НС і здоров'я людини.
	Макулатура	II	**Цинк** - 0,000053 – 0,000056 **Zn** **Свинець** - 0,000049 – 0,000051 **Pb** **Хром** - 0,000051 – 0,000054 **Cr** **Мідь** - 0,000033 – 0,000035 **Cu** **Целюл**	**Цинк** $T_{кип.}$= 913°С $T_{плав.}$= 4,19°С **Свинець** $T_{кип.}$= 1751°С $T_{плав.}$= 327,3°С **Хром** $T_{кип.}$= 1890°С $T_{плав.}$= 2480°С **Мідь** $T_{кип.}$= 2580°С $T_{плав.}$= 1083°С **Целюлоза** $T_{возг.}$ с обуглив. ≥	Негативний вплив на ОС і людини визначається його хімічним складом. **Цинк** Малотоксичний для теплокровних тварин при надходженні з їжею і питної водою-концентрація в питній воді 11,2 ... 26,6 мг / л переноситься без будь-яких ознак інтоксикації. Дуже корисний для флори, будучи одним з найважливіших мікроелементів харчування, однак лише в концентрації до 0,2 мг / л, крім того, елемент силяється до кумуляції в грантах. Дуже токсичний для водних організмів, порушуючи процеси самоочищення водойм і стаючи токсичним для іхтіофауни в концентрації 0,15 ... 5,0 мг / л. Мутагенна і онкогенна небезпека. **Свинець** У природних водах міститься в концентрації 0,001 - 0,023 мг / л. У концентрації 2,0 мг / л надає воді металевий присмак. Можливо має мутагенну і канцерогенну дію, значно збільшує токсичну дію інших металів. В конценрації 1,90 мг / л згубно діє на дафній, концентрація 0,1 мг / л погіршує процеси самоочищення водойм. Свинець токсичний для рослин в концентрації понад 5,0 мг / кг ґрунту. Помірно токсичний. Викликає хронічне отруєння. Має здатність вражати центральну і периферичну нервову систему, кістковий мозок і кров, судини, синтез білка, генетичний апарат клітини. **Хром** Міститься в природних водах в концентрації 0,001 ... 0,112 мг / л. LK50 Cr (VI) для риб-30,0 ... 50,0 мг / л, LK50 Cr (III) для риб- 117,0 мг / л. Низькі концентрації хрому позитивно впливають на ріст рослин, проте полив водою

			оза - 97,299 814 – 96,999 804 **($C_6H_{10}O_5)_n$** **Вода** - 2,7 – 3,0	100°С	С / Г культур з концентрацією хрому 10,0 ... 50,0 мг / л гальмує їх розвиток. На тварин надає загально токсичне, подразнююче, кумулятивне, алергенну, канцерогенну і мутагенну дію. Володіє канцерогенною властивістю. **Мідь** У природних водах міститься в концентраціях 0,001 ... 0,98 мг / л. У концентрації 0,5 мг / л забарвлює воду, в концентрації> 1,0 мг / л-помітно збільшує мутність води. Дуже токсична як для водних організмів, так і для рослин. У концентрації 0,001 мг / л гальмує розвиток синьо зелених водоростей, LK50 практично для всіх видів риб становить 0,18 ... 1,35 мг / л (короп, карась, окунь, щука, сом). Куммулюється ґрунтом і рослин-ями. У концентрації 0,1 ... 0,2 мг / л надає токсичну дію на ріст рослин. Високотоксичний метал. Викликає гостре отруєний-ня, має широкий спектр токсичної дії. **Целюлоза** Нетоксична. Досить легко підвержен біодеструкції лігнін- і целюлозоруйнуючими бактеріями і деякими класами низших грибів. У зв'язку з нетоксичністю LD50 для тваринах не установлена. Токсичність визначається за вмістом важких металів, здатних мігрувати з неї в навколишнє середовище. При попаданні на ґрунт, в воду і атмосферне повітря чинить негативний вплив на ОС і здоров'я людини.
	Побутові відходи	IV	**Побутові відходи** - 100 – 100, в т. ч.: **Папір** -30 - 17; [**($C_6H_{10}O_5)_n$** - целюлоза] **Поліетилен** -20 –	**Целюлоза** $T_{возг. с обуглив.} \geq$ 100°С **Поліетилен** - $T_{размяг.} \geq$ 150°С	Негативний вплив на ОС і людини визначається його хімічним складом. **Целюлоза** Нетоксична. Досить легко піддавав біодеструкції лігнін- і целюлозоруйнуючими бактеріями і деякими класами низших грибів. У зв'язку з нетоксичністю LD50 для тваринах не установлена. Токсичність визначається за вмістом важких металів, здатних мігрувати з неї в навколишнє середовище **Поліетилен** Нетоксичний для всіх видів флори і фауни в зв'язку з дуже високою біологічною інертністю. Нерозчинний у водних середовищах і не впливає на санітарний режим водойм. Використання його не вимагає запобіжних заходів. Отруєння можливі при виробництві та переробці плівки, в результаті виділення окису

			24; **(- CH_2 - CH_2 -)$_n$** **Харчові відходи** - 37 –44;	**Харчові відходи** $T_{биоразл.}$ ≥ 4º С	вуглецю, альдегідів, органічних кислот [66]. **Харчові відходи** Нетоксичні.

ДОДАТОК Ґ

Слайди презентації

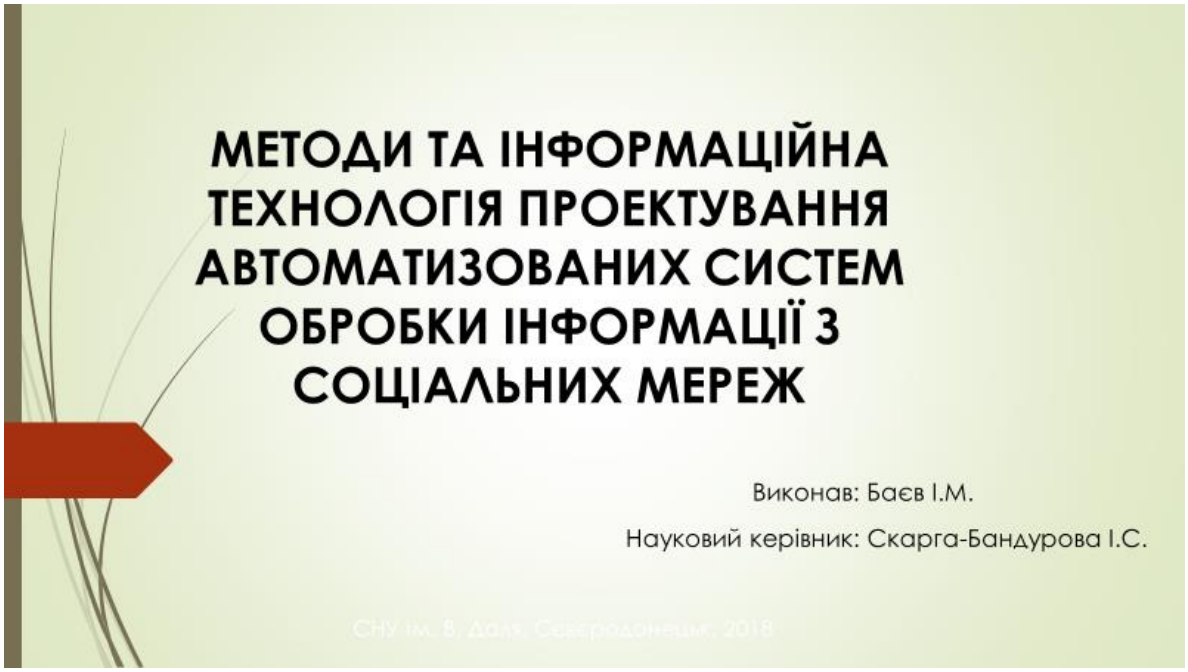

Рисунок Ґ.1 – Слайд №1

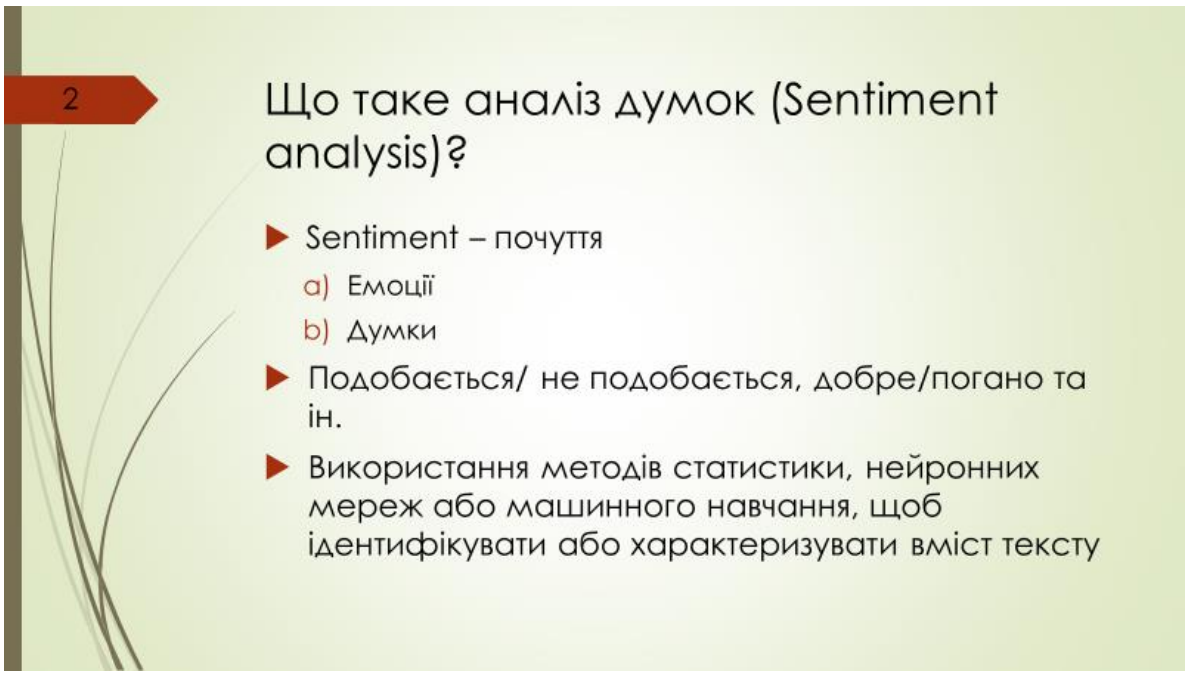

Рисунок Ґ.2 – Слайд №2

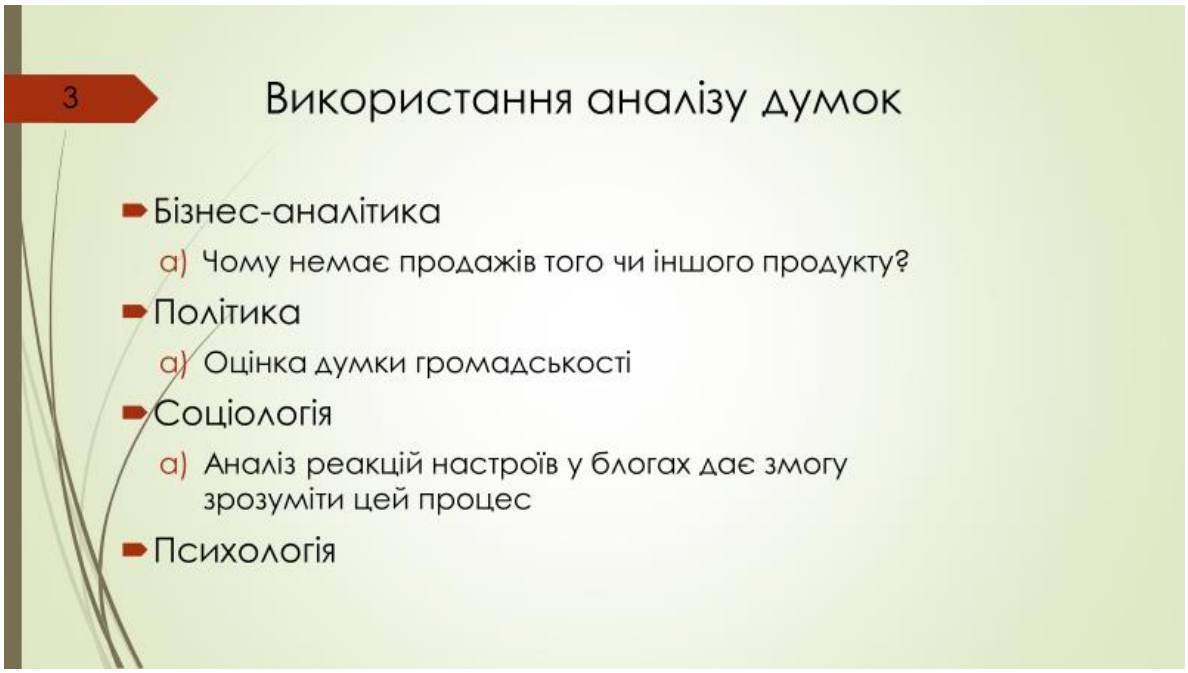

Рисунок Ґ.3 – Слайд №3

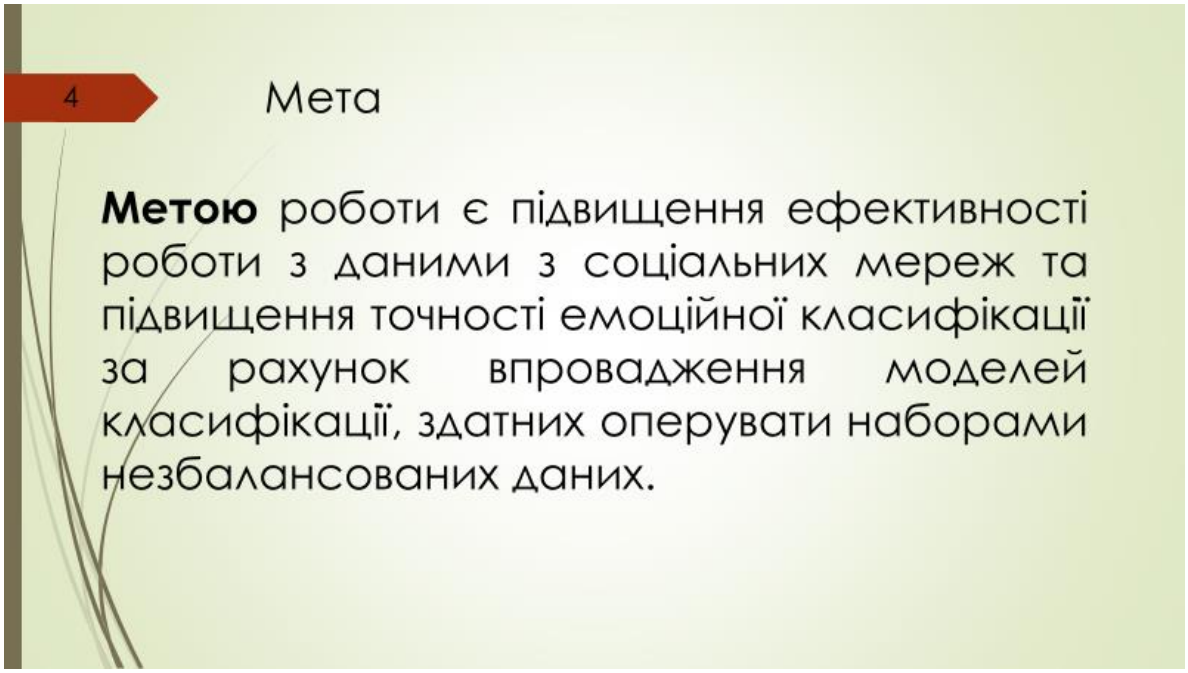

Рисунок Ґ.4 – Слайд №4

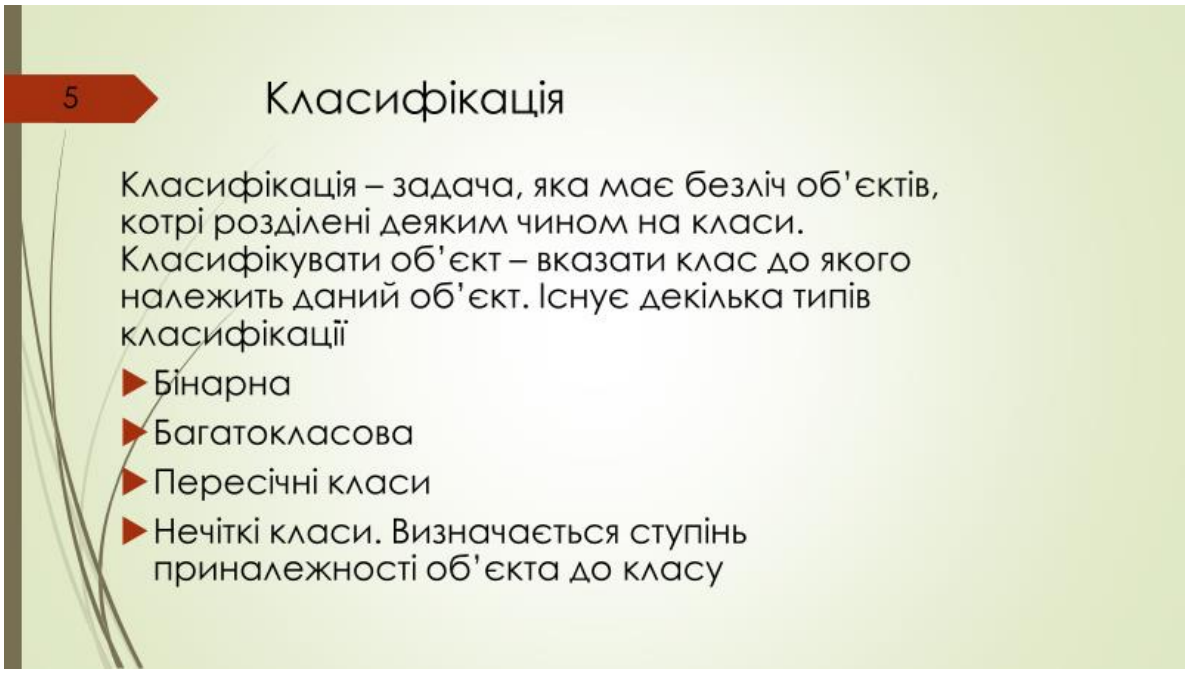

Рисунок Ґ.5 – Слайд №5

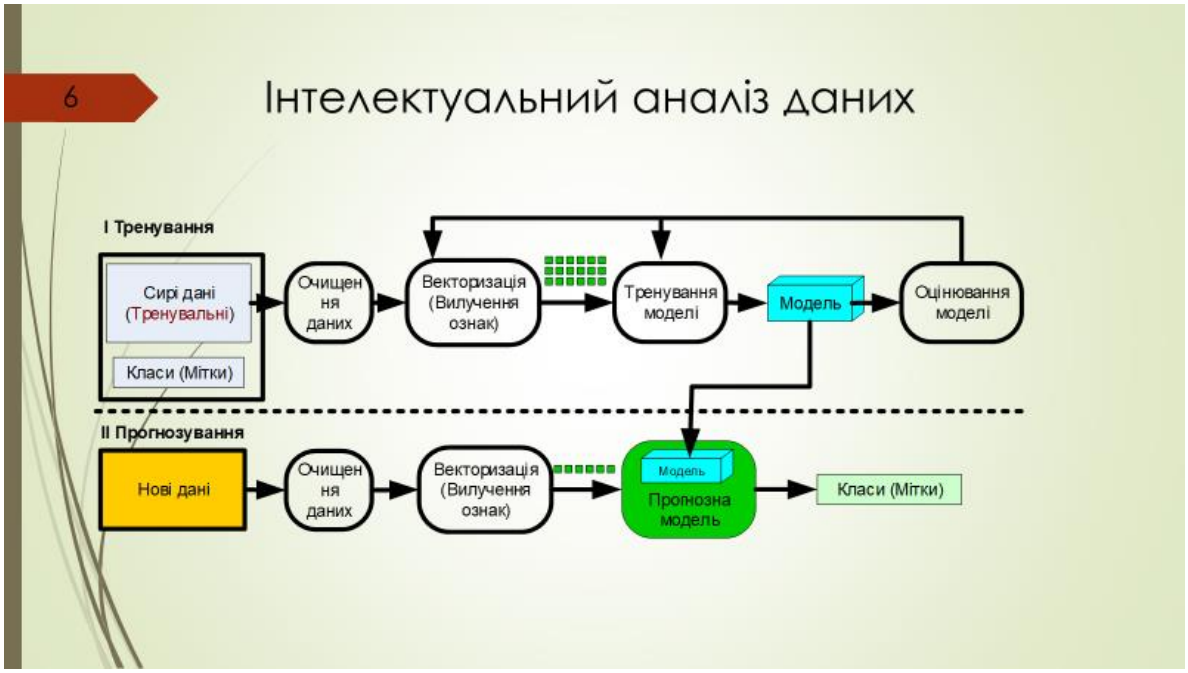

Рисунок Ґ.6 – Слайд №6

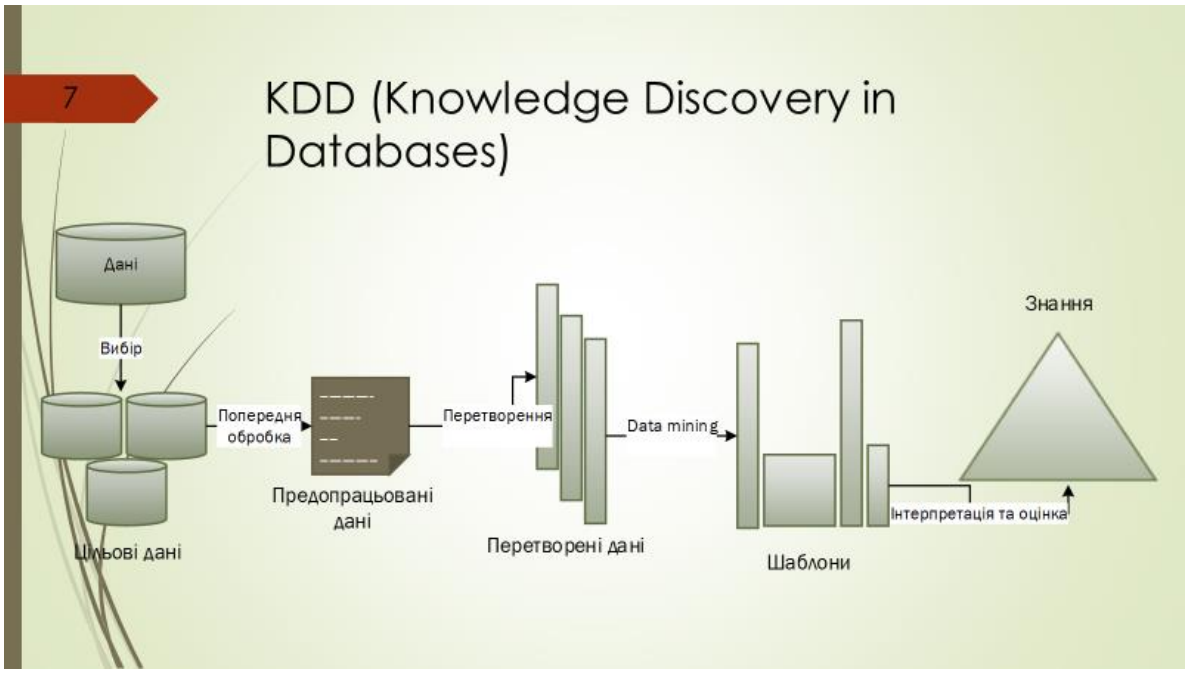

Рисунок Ґ.7 – Слайд №7

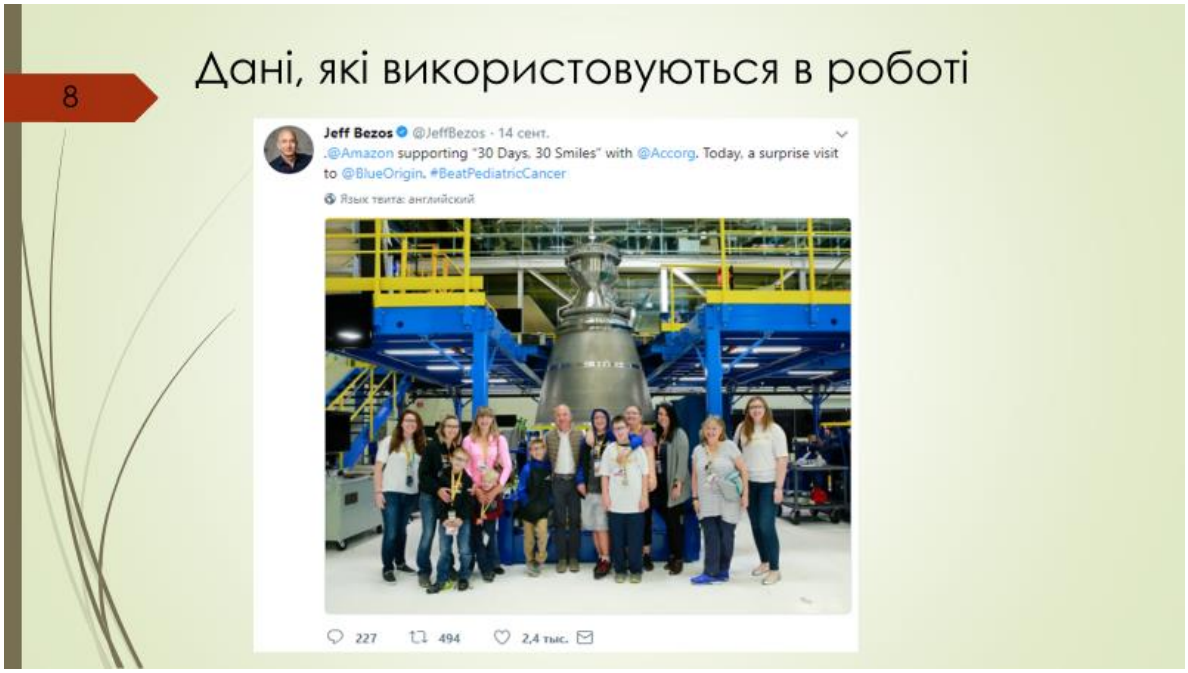

Рисунок Ґ.8 – Слайд №8

Рисунок Ґ.9 – Слайд №9

Рисунок Ґ.10 – Слайд №10

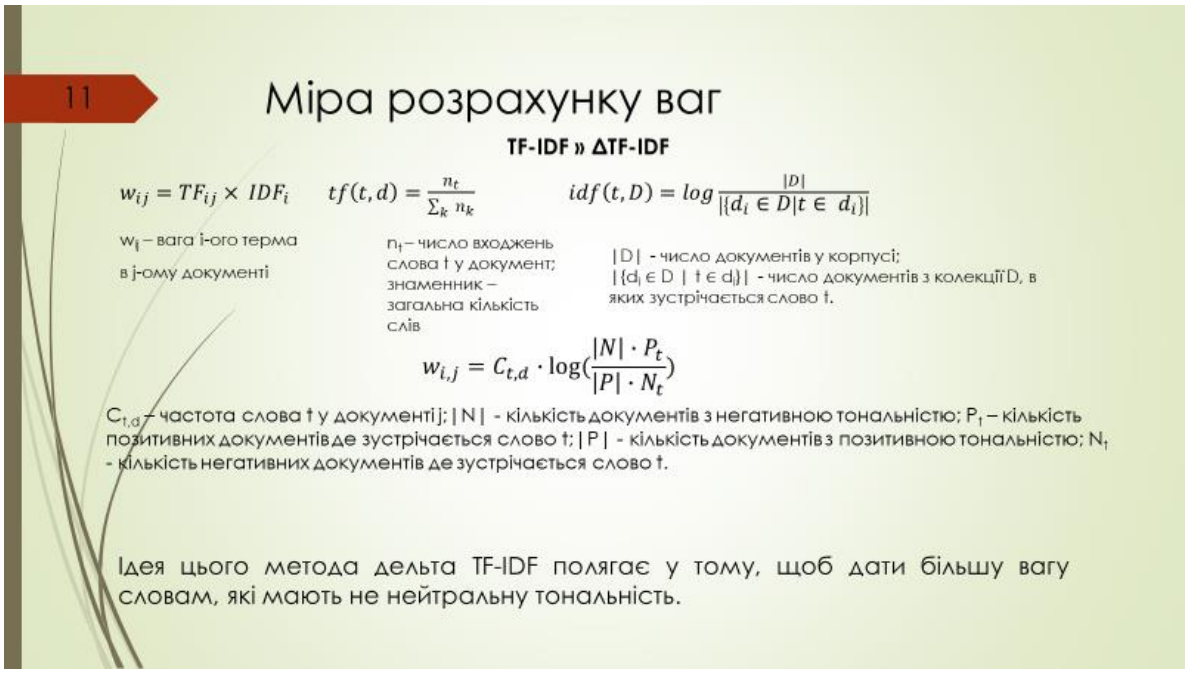

Рисунок Ґ.11 – Слайд №11

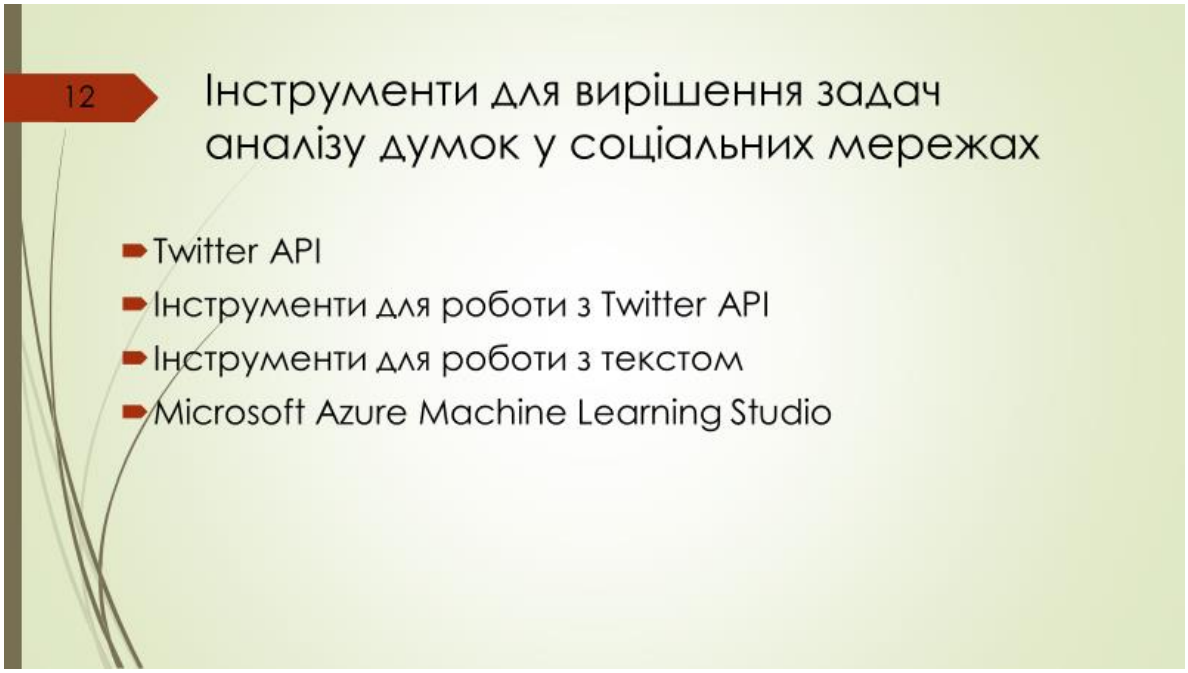

Рисунок Ґ.12 – Слайд №12

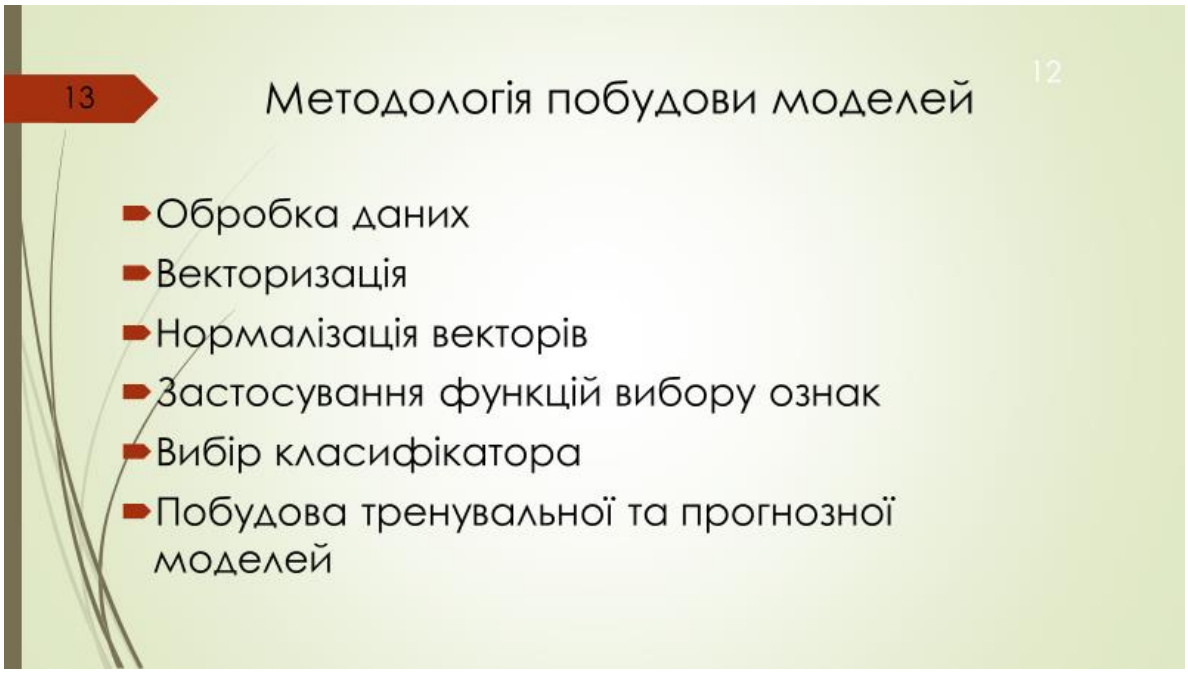

Рисунок Ґ.13 – Слайд №13

Рисунок Ґ.14 – Слайд №14

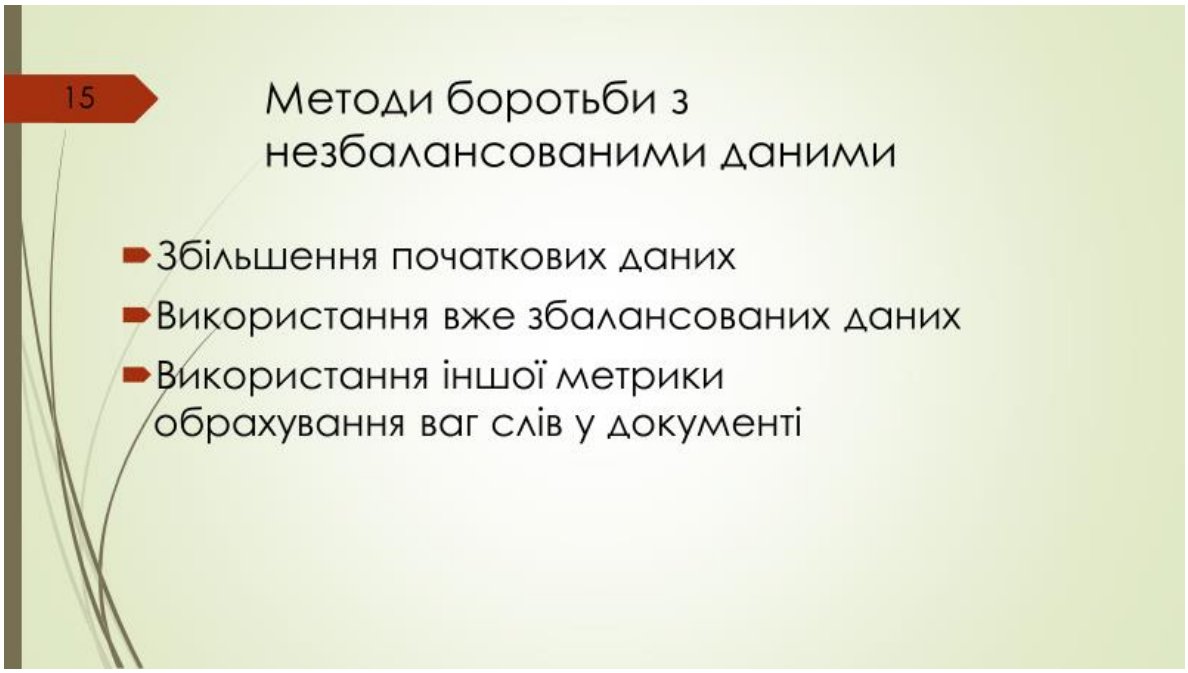

Рисунок Ґ.15 – Слайд №15

Рисунок Ґ.16 – Слайд №16

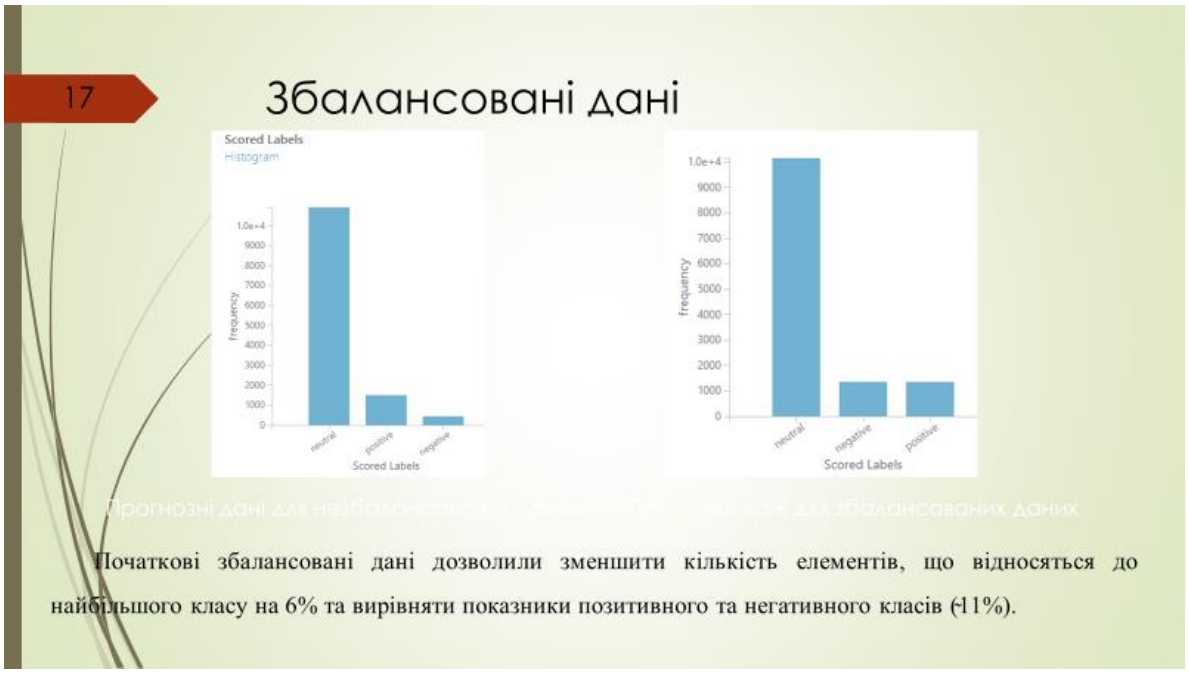

Рисунок Ґ.17 – Слайд №17

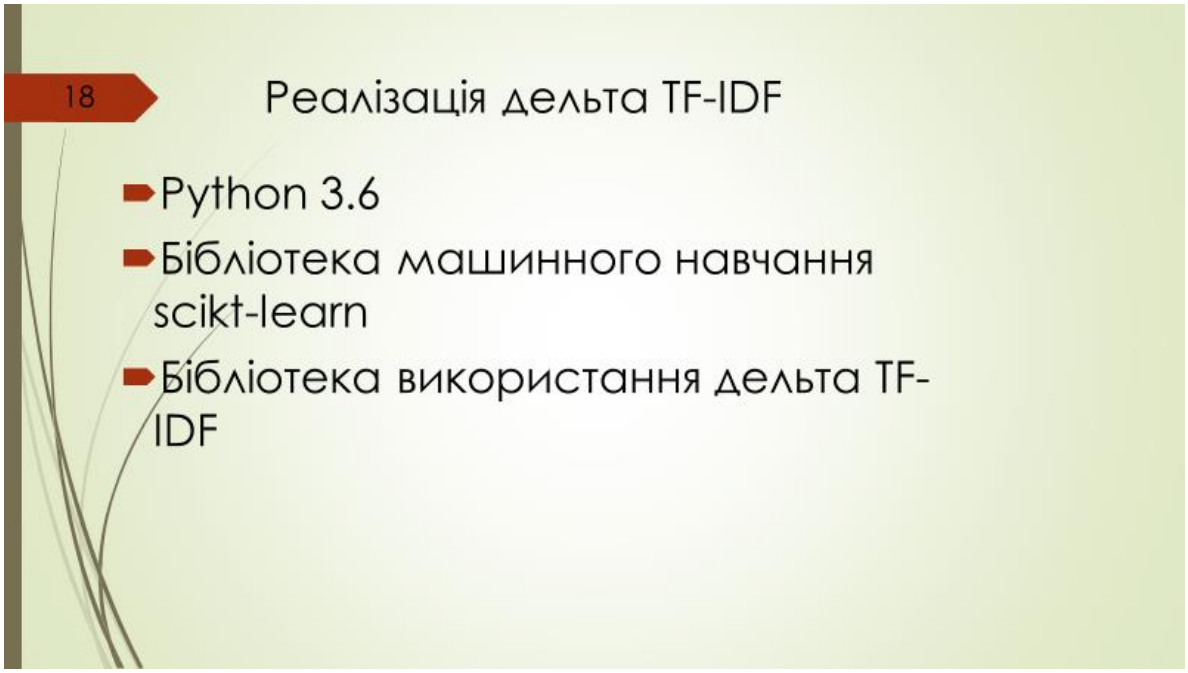

Рисунок Ґ.18 – Слайд №18

Рисунок Ґ.19 – Слайд №19

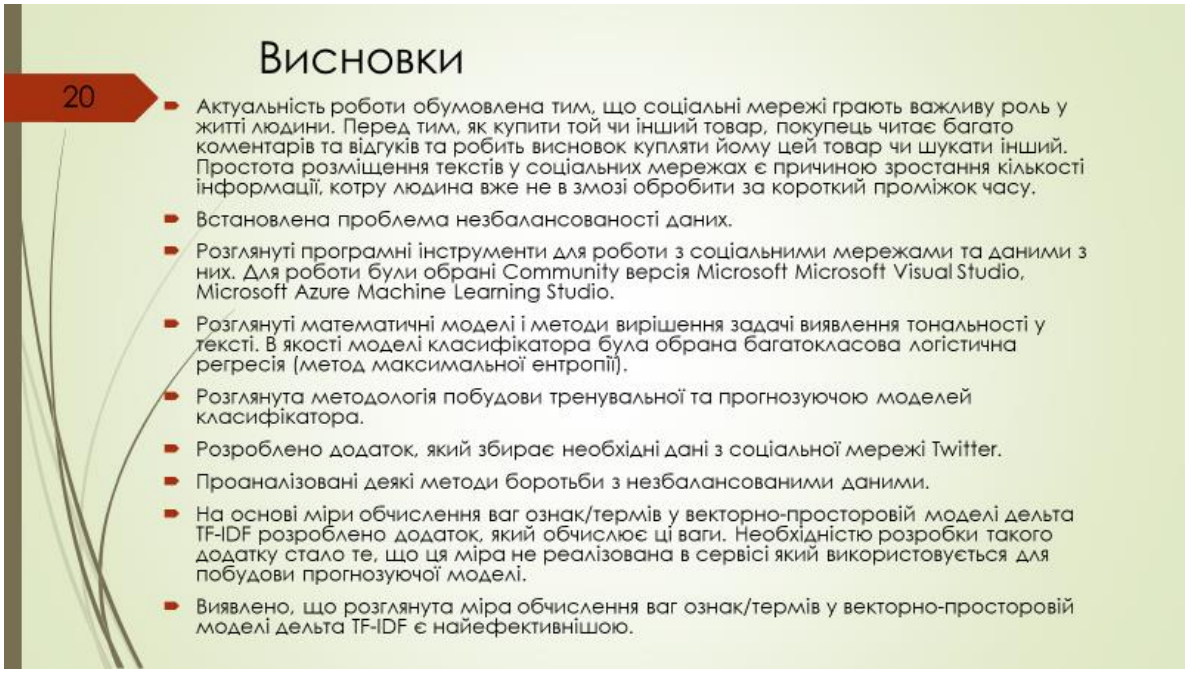

Рисунок Ґ.20 – Слайд №20